27656

National Library of Canada
Bibliothèque nationale du Canada

CANADIAN THESES ON MICROFICHE

THÈSES CANADIENNES SUR MICROFICHE

NAME OF AUTHOR/*NOM DE L'AUTEUR* Allan R. Hammond

TITLE OF THESIS/*TITRE DE LA THÈSE* Assessment Context in Relation to Physics Achievement and Cognitive Style

UNIVERSITY/*UNIVERSITÉ* The University of Alberta

DEGREE FOR WHICH THESIS WAS PRESENTED/ *GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE* Ph. D.

YEAR THIS DEGREE CONFERRED/*ANNÉE D'OBTENTION DE CE GRADE* 1976

NAME OF SUPERVISOR/*NOM DU DIRECTEUR DE THÈSE* Dr. H. Kass

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/*DATÉ* Nov. 26, 1975 SIGNED/*SIGNÉ* Allan R Hammond

PERMANENT ADDRESS/*RÉSIDENCE FIXE*
755 Irvine St.
Fredericton, N.B.
E3A 3E6

NL-91 (3-74)

INFORMATION TO USERS

THIS DISSERTATION HAS BEEN
MICROFILMED EXACTLY AS RECEIVED

This copy was produced from a microfiche copy of the original document. The quality of the copy is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Canadian Theses Division
Cataloguing Branch
National Library of Canada
Ottawa, Canada K1A 0N4

AVIS AUX USAGERS

LA THESE A ETE MICROFILMEE
TELLE QUE NOUS L'AVONS RECUE

Cette copie a été faite à partir d'une microfiche du document original. La qualité de la copie dépend grandement de la qualité de la thèse soumise pour le microfimage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

NOTA BENE: La qualité d'impression de certaines pages peut laisser à désirer. Microfilmée telle que nous l'avons reçue.

Division des thèses canadiennes
Direction du catalogage
Bibliothèque nationale du Canada
Ottawa, Canada K1A 0N4

THE UNIVERSITY OF ALBERTA

ASSESSMENT CONTEXT IN RELATION
TO PHYSICS ACHIEVEMENT
AND COGNITIVE STYLE

by

© ALLAN R. HAMMOND

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF SECONDARY EDUCATION

EDMONTON, ALBERTA
SPRING, 1976

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled Assessment Context in Relation to Physics Achievement and Cognitive Style submitted by Allan R. Hammond in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

H. Ross
Supervisor

Brown

Thomas E. Kieren

M. A. Nay

J. Kent Davis
External Examiner

Date Nov 25, 1975

ABSTRACT

The purpose of the investigation was to appraise the roles of cognitive style and physics achievement in relation to certain properties of tests used to assess physics achievement.

The assessment context of a test item was defined as a property indicating how closely the item conformed to the features of real-life, out-of-school situations as opposed to the features of standard classroom test situations. Physics items low in assessment context possess the features resembling the real-life situation, namely: 1) minimal redundancy of the essential physics information, 2) irrelevant physics information, and 3) two or more physics concepts among the response alternatives. The items high in assessment context possess: 1) redundancy of the essential physics information, 2) no irrelevant physics information, and 3) only one physics concept among the response alternatives. A 20-item, multiple choice test with items high in assessment context was constructed and two 20-item tests having items low in assessment context. One of the latter tests was a multiple choice test, the other consisted of diagrammatic items with a free response format.

The literature was reviewed on evaluating physics achievement and the cognitive styles of field independence and breadth of categorization. From the literature review and an analysis of the properties of the assessment context tests a number of hypotheses were constructed. An empirical study was conducted with six classes of high school physics seniors in order to test the hypotheses. The relationships among the variables were investigated by various statistical techniques including Pearson correlation, canonical correlation and multiple linear regression analysis.

The results of the investigation indicated that breadth of categorization was significantly positively associated with verbal and quantitative ability but not with physics achievement or assessment context. Field independence although unrelated to verbal and quantitative ability was significantly positively related with all measures involving physics achievement; however, when added to physics achievement as a predictor of success on the assessment context tests, field independence improved significantly the prediction of only the test consisting of items high in assessment context.

The study also produced evidence on the effects of certain non-standard features of test items and on the relationship of field independence with verbal and quantitative ability.

Arising from the findings of the study were a number of suggestions for further research and implications for the teaching of physics.

ACKNOWLEDGEMENT

The study could not have been completed without the assistance of numerous persons. I wish to express my appreciation to the following:

Dr. Heidi Kass for finding time in a busy schedule to offer extensive, helpful criticisms during the period of the study.

The Tests and Examinations Branch of the Alberta Department of Education for making available needed test results.

The Edmonton Public and Edmonton Separate School Boards for permitting access to the schools from which the sample was drawn.

The physics teachers and students at the three Edmonton high schools who participated in the study.

The professors and students of the five classes of Physics 100, The University of Alberta, who wrote the tests during the developmental stages.

The persons in the Division of Educational Research Services, The University of Alberta, who provided computer programs and offered advice on the analysis of data.

The members of the dissertation committee.

TABLE OF CONTENTS

CHAPTER PAGE

1. THE PROBLEM 1
    - Introduction 1
    - Purpose of the Study 2
    - Assessment Context 4
    - Cognitive Styles 16
    - Statement of the Problem 18
    - The Need for the Study 18
    - Design of the Study 21
    - Hypotheses of the Study 22
    - Definitions 26
    - Assumptions 30
    - Delimitations 30
    - Limitations 31
    - Summary 32
    - Overview 32

2. REVIEW OF THE LITERATURE 34
    - Field Independence 34
        - Foundations 35
            - Verbal abilities 37
            - Mathematical abilities 40
            - Sex differences 41
        - School-Related Research 43
        - Measurement 50
        - Instruments 50

Summary 60

Breadth of Categorization 63

Pettigrew Category Width Scale 64

Sorting and Categorizing 67

Risk-Taking 70

Sex Differences 71

School-Related Research 72

Mathematical and other abilities 73

Originality 75

Summary 78

Evaluating Physics Achievement 80

Applications from Concrete Situations 81

Generalized Basis for Deriving Assessment Items 82

Summary 87

3. INSTRUMENTATION AND DESIGN 89

Assessment Context 90

The LCT and HCT 100

Research related to the HCT and LCT 106

The DDT 108

Review of studies related to the DDT 109

Cognitive Style Variables 113

The HFT 113

The CWS 114

Ability Variables 116

SCAT-V and -Q 116

Physics Achievement Variables 118

The ASK and ASAT subtests 119

Design 121

The Sample 121

Test Administration 122

Mathematical Procedures 124

4. HYPOTHESES AND DATA ANALYSIS PROCEDURES 126

Hypotheses 126

Hypotheses of Overall Relationships 127

The Criterion Variables 129

Predictors of Criterion 130

The Covariates 130

Physics achievement 131

Breadth of Categorization 133

Field independence 137

Level of Significance 144

Data Analysis Procedures 144

Canonical Correlation 145

Interpretation 148

Multiple Regression Analysis 152

5. RESULTS AND DISCUSSION 156

Test Results 156

Preliminary results of physics achievement item ratings 156

Missing data 158

Means, standard deviations and reliabilities 161

Intercorrelations 163

Tests of Multivariate Hypotheses 167

Relationships Among Criterion Variables 174
Multiple Linear Regression Tests of Hypotheses 177
Verbal and quantitative ability 178
Tests with the HCT as criterion 178
Tests with the LCT as criterion 183
Tests with the DDT as criterion 192
Exploratory Analysis with the Assessment Context Tests as criteria 195
Summary of the Results and Discussion 199
6. CONCLUSIONS AND RECOMMENDATIONS 205
Summary 205
Purpose 205
Major hypotheses 206
Method 208
Results 209
Conclusions 214
Implications for Science Education 214
Implications for Further Research 219
Limitations of the Study 221
FOOTNOTES 225
REFERENCES 226
APPENDIX A Instructions for Rating Assessment Context 236
APPENDIX B The Physics Test 246
APPENDIX C The Discrepancy 260
APPENDIX D Raw Scores 267

APPENDIX E Additional Tables 273
APPENDIX F The Category Width Scale 275

LIST OF TABLES

TABLE PAGE

1. Summary of the Evidence for Practice Effects in the Mean Scores on the HFT 59
2. Summary of the Pretest Data for the PT which Includes the HCT and LCT as Sub-tests 104
3. Summary of the Pretest Data for the DDT 110
4. The Sample by School and by Sex 123
5. Missing Data in the Sample of 139 Subjects 159
6. Summary of Test Results 162
7. Matrix of Intercorrelations of Test Scores 164
8. Canonical Correlation Coefficients and Bartlett's Significance Test Results 169
9. Structure Coefficients for Criterion and Predictor Variables for the First and Second Canonical Correlations 170
10. Tests of Hypotheses on changes in the Multiple Correlation of the HCT as the Criterion and Various Predictors 181
11. Tests of Hypotheses on Changes in the Multiple Correlation of the LCT as the Criterion and Various Predictors 185
12. Tests of Hypotheses on changes in the Multiple Correlation of the DDT as the Criterion and Various Predictors 193
13. Normalized Weights Applied to the Criterion and Predictor Variables for the First and Second Canonical Correlations 274

LIST OF FIGURES

FIGURE PAGE

1. A Diagrammatic Item from the Field of mechanics 14

2. A Diagrammatic Item from the Field of electrostatics 98

CHAPTER 1

THE PROBLEM

Introduction

That students be able to apply their science knowledge beyond the classroom has been a continuing concern of science teachers (Smith and Tyler, 1942; Lewis, 1965). However, the out-of-school, or real-life situations in which science knowledge can be applied are difficult to analyze from the educational point of view since they are complex as compared with problem situations described in textbooks. If a way were found to capture the essence of real-life problems so as to permit their presentation in classrooms, then teaching-learning procedures could be devised to enable students to solve such problems and methods of assessment implemented to determine if the procedures were successful. To say that real-life situations must be analyzed, teaching-learning procedures devised and assessment methods implemented is not to say that schools are failing at the present time to prepare students. Students are applying much of what they have learned, and further analysis and efforts may not result in further improvement. Yet, the continued stress on the application-to-life objective implies that all is not as it could be. Hurd (1973) has encouraged science teachers to improve their efforts toward educating students who are able to attack the science-social problems facing them today. The present investigation is an attempt to elucidate some of the factors involved in preparing science students who are able to apply their knowledge in out-of-school situations. The particular focus is on assessing students' knowledge of physics concepts by methods having some features of the real-life situation.

Purpose of the Study

In raising the issue of the capability of students for applying their knowledge in real life situations several questions emerge: Can students be taught in ways which are highly effective to secure the objective? Are they presently being taught in ways which are effective? How can assessments of the objective be made in the school situation? The present study is not concerned with the teaching methods nor the content of the science courses as such but is concerned with assessment of the real-life objective in the school situation.

In comparing and contrasting the real-life application of knowledge and the classroom assessment of application of knowledge several differences become apparent. For example, more irrelevant information is present in the real-life situation than in the classroom assessment situation. The approach followed in the present investigation is to modify certain aspects of the classroom assessment situation so that some qualities of the real-life situation are incorporated. The basis of the modifications is outlined in a later section of the chapter at hand, and the specially constructed physics tests are described in detail in Chapter 3. One of the main questions of the study is: What is the relationship of student achievement in physics in the modified and unmodified assessment situations?

The nature of the modified achievement tests suggests that perceptual or cognitive abilities in addition to science knowledge may be relevant in student performance on the tests. The second main question of the study is: Are other abilities besides science knowledge related to achievement on the modified and unmodified tests? Further explanations giving more detail on the procedures followed are presented in the paragraphs which follow.

The carrying out of a logical analysis of real-life situations in relation to applying science knowledge was the first step of the present investigation. The purpose of the analysis was to ascertain the factors with which a person must contend in applying what has been learned in science to a real-life situation in which a problem exists. The factors which emerged from the analysis were used to guide the preparation of test items having characteristics similar to the real-life situation. The extent to which the characteristics were found in test items was defined as related to the assessment context of the items. The nature of the resulting test items suggested that certain characteristic ways of perceiving, or cognitive styles of students, might have an important bearing on their capabilities for answering the items. Hence an investigation of the relationship of assessment context and cognitive style was undertaken. It also seemed important to examine the relationships of assessment context and cognitive style with respect to other variables with which science teachers are concerned, namely verbal and quantitative ability and science achievement. Because science achievement and cognitive style are such broad concepts the study had to be narrowed for manageability; hence the physics achievement of high school seniors and the two cognitive style dimensions of extent of field independence and breadth of categorization were chosen. The reasons for the particular choices are described in appropriate subsections below.

The present chapter continues with explanations of the meaning of assessment context and of the cognitive style variables. After clarifying the purpose and need for the study the major hypotheses and design are presented. The chapter concludes with statements of definitions, delimitations and assumptions followed by an overview of the entire study.

Assessment Context

Assessment context, as applied in the present study, is a variable property of test items. Assessment context is described more fully below and its characteristics are shown to be derived from the attributes of those actual physical situations which exist in our environment and to which a knowledge of science may be applied. Furthermore the characteristics of the assessment context may be related to concepts in psychology and science education.

Assessment context is a property of test items indicating how closely the test items conform to certain features of a real-life, out-of-school problem situation. The characteristics of the real-life situations to which the assessment property pertains include the degree to which the intent of the problem is defined, the extent of "noise" or irrelevant data present, and the number of alternative solutions available. To illustrate more specifically the characteristics of the real-life situation as opposed to the classroom test situation examples of each are presented below and salient features, in relation to assessment context, are discussed.

Suppose that a student has been studying in physics about temperature differences and the methods of heat transfer. He is presented with the following test item:

> You are required to bring to a boil a pot of water in as short a time as possible, given an electric stove, pot and lid, some water. State how you would heat quickly the water. Be sure to mention the roles played by conduction, convection and radiation.

Now consider the same student at home in a situation where he wants to boil some water quickly to cook a weiner. Let us compare his performance on the test item with his performance in the real situation with the assumption that he knows the physics he has studied.

In both situations what is required has been made clear to him, one way or another, so the problems are the same in that respect but they differ in others. With the test item the principles of physics which he must apply are suggested to him, namely those related to conduction, convection and radiation. In the actual situation, however, the words are not written for him to read so he does not have the particular direction of his attention and the reminders or cues which he had in the test item. Furthermore, the actual situation contains much irrelevant information not in the test item. For example, he may have available a large pot, or a small pot, the stove may have a large element or a small element, warm or cold water may be available to put in the pot, and one pot may have a tight fitting lid and another a loose one. The actual situation is seen to contain more information than the test item and some of the information in the actual situation is irrelevant insofar as the verbal test item is concerned.

Hedges (1966, p. 10) has noted that a good science test item should have a central theme. In a real-life situation, however, there are other factors to consider as well as the main question. In the example above, the type of pot to choose, or the heating element on the stove to use are questions which arise and which may distract from the main theme. Ebel (1972, p. 199) has stated that good test items should not contain unnecessary or irrelevant material, yet real-life situations do. In the above example the weiners of the real-life situation may or may not have plastic casings. This datum is irrelevant yet it may be potentially distracting to some students. The question then arises: Is it possible to construct classroom test items which contain representations of real-life situations not normally included in good test items as defined by Ebel (1972) and Hedges (1966)? The answer seems to be yes, and the dimension along which test items vary in their resemblance to the real-life situation as opposed to the "good" test item of the classroom is defined, in the present study, as the assessment context of items. Items closely resembling the test items of the classroom are said to be high in assessment context, or high context items; those resembling the real-life situation are said to be low in assessment context. The assessment context dimension seems to be a directiveness-nondirectiveness dimension with classroom items decidedly more directive than the real-life situations. Items low in assessment context are therefore considerably more nondirective than high context items. The properties of test items on which the assessment context depends are: 1) the extent of the directiveness of the information leading to the correct response, and 2) the amount of information which is irrelevant to making the correct response to the item.

The extent of the directiveness of the information in a test item leading to the correct response is closely related to the redundancy of the essential information. Hunt (1962) has defined information as being redundant if it is consistent with all hypotheses currently held. A physics test item may exhibit in several ways redundancy of the information needed for a correct response. The item may name the required science principle and in addition specify the conditions of its operation. For the student who knows the principle well the conditions of operation are recalled on the reading of the name of the principle and hence the further statement of the conditions in the item represents redundant information for the student since the stated conditions are consistent with his recall of the conditions. His hypothesis on the meaning of the physics principle continues to be held and is confirmed. For the student who is uncertain in his understanding of physics the redundant information contains two methods for him to arrive at the correct basis of solution. If he does not recall the meaning of the physics principle from its name, he may be able to recollect the meaning from the conditions of operation of the principle which are also stated.

Besides naming principles as well as stating the conditions of operation of the principles, test items can exhibit redundancy of essential information in other ways. The naming of physical quantities as well as specifying their units of measurement is another example of redundancy, as is the naming of physical quantities in addition to providing their standard abbreviations. Redundancy in the information that is essential for making the correct response tends to reinforce any tendency the student may have toward the correct response. Hence such redundancy tends to make the item high in assessment context. Items low in assessment context do not exhibit redundancy of physics information, hence they are less directive toward the required response.

The directiveness - nondirectiveness of a test item is also related to the presence of irrelevant information in the item. That real-life situations tend to contain irrelevant information has been noted above, and this property tends to make them nondirective to the intended response and hence low in assessment context.

Methods of incorporating irrelevant yet plausible information into a test item are to state unnecessary additional facts about the physical properties of objects or to list other conditions which are not essential. In multiple choice items irrelevant information can be readily added by not having different physics concepts among the various response alternatives. Items containing no irrelevant information are higher in assessment context than items which contain such information. The characteristics which distinguish between items high and low in assessment context are described and illustrated in greater detail in Chapter 3 where the contruction of test items is explained.

The two generalizations discussed above on the differences between conventional classroom test items which tend to be high in assessment context and real-life situations which are low in assessment context lead to three criteria to be used in distinguishing between them:

1. Low context items do not have the redundancy of essential information exhibited by high context items
2. Low context items contain irrelevant yet plausible scientific information
3. Low context items contain several different physic concepts among the response alternatives while high context items have only one.

Presented below are examples of a high context item followed by a low context item both intended to measure the capability of calculating kinetic energy of an object.

Item high in assessment context:

A motionless box is free to move without friction. The box has a mass (M) of 10 kg. A constant force (F) of 20 newtons acts on the box through a distance (d) of 9.0 metres. After the force has acted the velocity (v) of the box is 6.0 m/sec. The kinetic energy (KE) of the box is:

A. 200 joules
B. 180 joules
C. 225 joules
D. 90 joules

Item low in assessment context:

A motionless box having a density of 40 units is located in a gravitational field of 10 units. The box is acted on by a net force of 20 units during which time the box moves 9 distance units. The coefficient of friction in effect as the box moves is 0.30. After the application of the force

A. the momentum of the box is 100 units
B. the force of friction is 3 units
C. the kinetic energy of the box is 180 units
D. the speed of the box is 180 units.

The high context item has redundant information in that the abbreviations and unit names for the various quantities as well as the usual names are presented. Furthermore, there are two alternative methods of arriving at the correct result either by taking the product of force and distance, or by finding the square of the velocity from the product of twice the force by the distance divided by the mass and then using the result of the calculation in the expression ($1/2\ mv^2$) to find the kinetic energy.

The low context item has no redundancies of the kind just described. In contrast, the high context item has no information which is not useful, or potentially useful for reaching the correct solution whereas the data in the low context item on the density, the gravitational field and the coefficient of friction are irrelevant to the solution of the problem. The response alternatives of the high context item have only one concept among them, that of the joule of energy. In contrast, the alternatives in the low context item contain the concepts of momentum, force, kinetic energy and speed.

The features of low context items which distinguish them from high context items also distinguish them in some respects from items written by the usual standards of good test items. Ebel (1972) in describing the requirements of well written test items, notes that items should not contain "window dressing" and that the item stem should specify what is required of the respondent, usually by having the verb of the sentence containing the response alternative located in the item stem and not in the response alternative. The irrelevant data in the item stems of the low-assessment context items as defined in the present study can be construed as window dressing, and the varied concepts of the response alternatives in these items usually require that the verb be placed not in the stem but in each response alternative. Thus certain features of the low context items mark them as being non-standard in Ebel's terms. Research related to non-standard item-writing practices suggests that low context items are not necessarily less effective even though they do not exhibit all of the features recommended by Ebel (1972). The non-standard nature of low context items is considered further in Chapter 3.

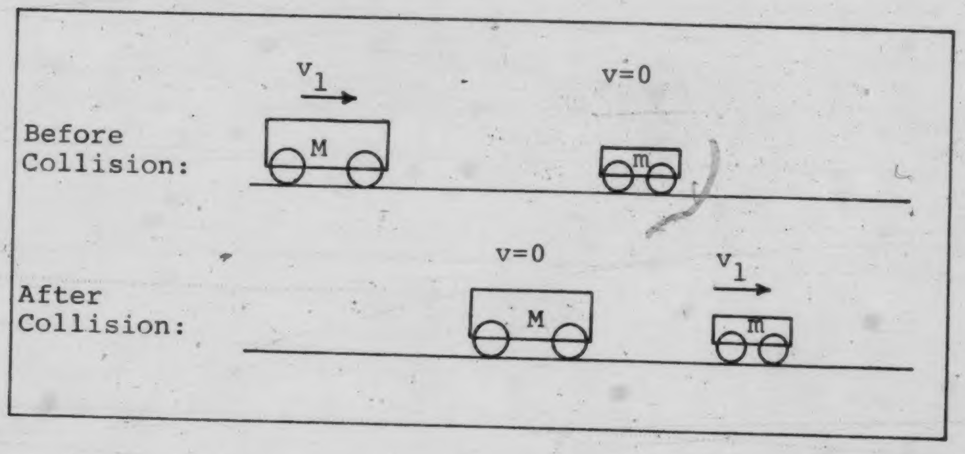
The properties of low context items presented thus far have been envisioned as applying mainly to verbal, multiple choice items. The properties also can be applied to a type of diagrammatic item having minimal verbal content and free-response format. Such items were also designed for use in the present study. The diagrammatic items present a situation to which the concepts and laws of physics may be applied. The respondent is asked to examine the situation for any features which seem to contradict what he would expect to find based upon his knowledge of physics. Because of the requirement to look for contradictions the test composed of diagrammatic items is referred to as the Discrepancy Detection Test. The respondent is informed that the diagrams may or may not contain potential contradictions. He is asked to indicate whether or not he perceives a contradiction and, if he finds such a discrepancy to explain the nature of it. The nature of the diagrams is such that there is minimal redundancy of the essential information, there is irrelevant data, and there exists the possibility of alternative physics concepts for explanation of apparent discrepancies. While discrepancy detection items and their relationship to low context items are discussed at length in Chapter 3, one discrepancy detection item is presented below in Figure 1.

Figure 1 does present a discrepancy. The law of

Figure 1

A Diagrammatic Item from the

Field of Mechanics



v_1
v=0
Before Collision:
M
m
v=0
v_1
After Collision:
M
m

physics violated is the law of conservation of momentum. The small cart should have a velocity less than v_1 after the collision and the large cart should still have a velocity to the right although the speed should be less than that of the small cart.

Examination of high context, low context and discrepancy detection items described above suggests that the processes used in responding to them may not be the same. With high context items the redundant information and the single concept response alternatives both tend to direct the student's attention toward a specific concept or principle of physics. Once this knowledge has been recalled the student applies it and then searches for the appropriate response among the alternatives presented. With low context verbal items and discrepancy detection items the student may follow a more complex process: picking out various pieces of self consistent data, assessing whether or not a potentially correct response is available, examining the data in another way to see if it is conformable with an hypothesis he may have arrived at from a given response alternative or from his analysis of a given diagram, or re-focusing on a different aspect of the data or diagram. The procedures followed by the student in arriving at the correct response in low context verbal and discrepancy detection items appear to be more complex than for the high context verbal items; that is, the information-processing requirements may be different for items which differ in assessment context.

Messick (1970, pp. 188-190) states that cognitive styles relate to the way a person customarily perceives, remembers, thinks, or solves problems and that cognitive styles are essentially information-processing habits. Hence differing stylistic tendencies among individuals may be related to differential capabilities in responding to test items high in assessment context as compared with those low in assessment context. The relevance of cognitive styles with respect to performance on high context verbal, low context verbal and discrepancy detection items is considered further in the next section.

Cognitive Styles

Of half a dozen or more stylistic dimensions which have been studied (Klein, 1967), two which may be particularly relevant in responding to physics test items are extent of field independence (Witkin, Dyk, Faterson, Goodenough and Karp, 1962) and breadth of categorization (Pettigrew, 1958). Witkin et al. (1962) have described extent of field independence as related to the ability to perceive items as separate from their backgrounds, or in general, to be able to overcome the influence of an embedding context. Furthermore, Witkin et al. (1962) have noted positive relationships between perceptual and intellectual tasks having in common the requirement of overcoming embeddedness. Individuals who are highly field independent are capable of overcoming embedding contexts. The low context test items and the discrepancy detection items of the present study may favor the more field independent student. The selecting of specific data from item stems or diagrams, the formulating and reformulating of hypotheses based upon various response alternatives or perceptions of diagrams, are processes which seem positively related to a high degree of field independence. Since high context items tend to be more directive and to require less complex response processes, a high degree of field independence is not anticipated to be so important in responding to them as compared to items which are low in assessment context. Thus one of the major questions of the present study concerns the relationship of assessment context and field independence.

The second cognitive style dimension to be investigated in relation to performance on tests varying in assessment context is breadth of categorization (Pettigrew, 1958). Breadth of categorization refers to the range of stimuli or qualities that are assigned by an individual to a common category in a task requiring the individual to group together stimuli or qualities which seem to be the same (Bruner and Tajfel, 1961, p. 231). The preceding definition of breadth of categorization has been adopted for the present investigation. The broad categorizer accepts as the same or closely related a range of differing stimuli whether they be objects to be sorted according to various properties, or differing approximations as to the size of some quantity. The capability for answering low context and discrepancy detection items seems related to grouping together, or distinguishing among various verbal or visual stimuli. Hence breadth of categorization may be associated differentially with capabilities for answering correctly test items which are high and low in assessment context. In Chapter 4 detailed arguments are presented suggesting an hypothesis for the relationship of assessment context and breadth of categorization.

Statement of the Problem

The purpose of the study is to examine the relationships which exist between performance on physics test items varying in assessment context and on measures to assess two cognitive styles. The main questions are:

1) What is the relationship of assessment context and extent of field independence?

2) What is the relationship of assessment context and breadth of categorization?

3) How are physics achievement and verbal and numerical ability associated with assessment context and the two cognitive styles?

The Need for the Study

Hurd (1973) has recently urged teachers to improve their efforts at preparing students to apply science knowledge to the environmental problems they meet after leaving school. That Hurd should feel the need to encourage teachers toward this objective is somewhat surprising because teachers have always given the objective high priority. The science teachers of the Eight-Year Study gave it highest priority (Smith and Tyler, 1942, pp. 77-111), and Lewis (1965) traced the continuing high concern of British teachers for the objective over a 25-year period. Hurd's (1973) concern, however, is understandable since the application-to-life objective is difficult to achieve. Bloom (1965) has noted the difficulties involved in preparing objectives and test items which are adequate, and Nedelsky (1965) has stated that it is nearly impossible to deduce from any real-life problem the precise pattern of attitudes, knowledge and approaches required to solve it. There is, then, a need to find ways of preparing students who are better able to apply what they have learned in science. Several different attacks have been made on the problem.

The teachers of the Eight-Year Study (Smith and Tyler, 1942) tried to analyze the application objective more clearly and found that more illustrative examples based upon commonly occurring phenomena within the experience of students were required in the teaching of science. Rowe (1965) claims that most objects or concepts in science vary in their meaning according to the context in which they are contained. The concept "pencil" is entirely different when the object is used as a writing instrument and when it is used to prop up a window. Rowe (1965) advocates varying the learning context so as to maximize the generalizability, and hence the applicability of what was learned. The recommendation of Hurd (1973) that students should be taught multiple methods and multistep approaches seems similar to the approach of Rowe (1965). Whatever approaches may be taken in teaching for the applicability of knowledge there is a need to assess their effectiveness. The low context test items, based as they are on analysis of cognitive aspects of real-life situations, may be valuable in assessing the application objective.

The relationship of the assessment context variable and physics achievement needs to be established. In the present study physics achievement has been subdivided as physics achievement at the knowledge level or at the algorithmic thinking level of Avital and Shettleworth (1968). The subdivision permits a more detailed analysis of the associations of high context, low context, and discrepancy detection items with knowledge and algorithmic thinking than could be done if physics achievement were classified as a unitary concept. If the results of the present study are to be of value to the curriculum developer or to the teacher, assessment context must be understood in relation to the levels of physics achievement and also in relation to intellectual ability. Since verbal and quantitative ability are the two dimensions of intellectual ability which are most relevant in the school situation, they will be used as covariates in assessing the relationships between assessment context and physics achievement.

Messick (1970) has pointed out that cognitive style variables play unknown roles in classrooms. The present study includes an examination of the association of field independence and breadth of categorization with achievement on items which vary in assessment context; hence, the study may clarify the role of two cognitive style dimensions with respect to one aspect of classroom performance. The design employed for appraising the relationships of the cognitive style dimensions and the other variables of the study is presented in the following section.

Design of the Study

The present study is a correlational one examing the existing states of variables with the administration of tests taking place within a 2 1/2-week period.

The assessment context variable, the criterion variable of the study, was chosen because of its importance for furthering our understanding of the feasibility of the application objective in teaching. Physics was selected as the content field for preparing the assessment context tests because of the physics academic background and teaching experience of the investigator. The nature of the low context test and discrepancy detection items suggested that cognitive styles could play a role in making responses to the items. A review of studies on cognitive styles indicated that extent of field independence and breadth of categorization were two style dimensions likely to be pertinent to an individual's performance on the items.

Finally, any study of relationships in effect for school students needs to be placed in the context of intellectual ability and school achievement. In the present investigation physics achievement has been treated as a covariate of the relationship between the assessment context and the cognitive style variables.

The sample for the study comprised 144 grade 12 Physics 30 students from three Edmonton high schools. Tests for the criterion and independent variables were administered by the investigator and one physics teacher near the end of the 1971-72 school year. The ability tests were administered as part of the testing program of the Alberta Department of Education.

The relationships among the variables were assessed with the aid of various correlational techniques including multiple linear regression and canonical correlation. The hypotheses which were tested are presented in the section which follows.

Hypotheses of the Study

The development of the hypotheses is explained in detail in Chapter 4 of the report. There are hypotheses on the overall relationships among the variables, followed by hypotheses on interrelationships among the assessment context tests, and hypotheses of associations between each of the assessment context tests and physics achievement and the cognitive style variables.

1. Hypotheses concerned with the overall relationships among the variables:

   1.1 The first canonical correlation between the set of assessment context variables and the set of variables including cognitive style, intellectual ability and physics achievement is zero.

   1.2 The second canonical correlation between the set of assessment context variables and the set of variables including cognitive style, intellectual ability and physics achievement is zero.

2. The following hypotheses test the relationships among the assessment context tests:

   2.1 The partial correlation between the *High Context Test* (HCT) and the *Low Context Test* (LCT) is zero when the influence of physics achievement is eliminated.

   2.2 The partial correlation between the HCT and the *Discrepancy Detection Test* (DDT) is zero when the effects of physics achievement are removed.

3. The following hypotheses assess the impact of physics achievement on performances on the assessment context tests:

   3.1 There is no significant increase in the multiple correlation coefficient (R^2) between the HCT and the ability variables when physics achievement at the knowledge level is added to the set of predictor variables.

3.2 There is no significant increase in the R^2 between the LCT and the ability variables when physics achievement at the knowledge level is added to the set of predictor variables.

3.3 There is no significant increase in the R^2 between the DDT and the ability variables when physics achievement at the knowledge level is added to the set of predictor variables.

3.4 There is no significant increase in the R^2 between the HCT and the ability variables plus physics achievement at the algorithmic thinking level is added to the set of predictor variables.

3.5 There is no significant increase in the R^2 between the LCT and the ability variables plus physics achievement at the knowledge level when physics achievement at the algorithmic thinking level is added to the set of predictor variables.

3.6 There is no significant increase in the R^2 between the DDT and the ability variables plus physics achievement at the knowledge level when physics achievement at the algorithmic thinking level is added to the set of predictor variables.

4. The following hypotheses are designed to assess the contribution of breadth of categorization toward performances on the assessment context tests:

4.1 There is no significant increase in the R^2 between the HCT and the ability measures plus physics achievement when breadth of categorization is added to the set of predictor variables.

4.2 There is no significant increase in the R^2 between the LCT and the ability measures plus physics achievement when breadth of categorization is added to the set of predictor variables.

4.3 There is no significant increase in the R^2 between the DDT and the ability variables plus physics achievement when breadth of categorization is added to the set of predictor variables.

5. The following hypotheses are designed to assess the contribution of the field independence variable toward performances on the assessment context tests:

5.1 There is no significant increase in the R^2 between the HCT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

5.2 There is no significant increase in the R^2 between the LCT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

5.3 There is no significant increase in the R^2 between the DDT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

Definitions

The terms defined below have been categorized into groupings related to assessment context, cognitive style, physics achievement and intellectual ability.

The definition of terms related to assessment context are presented first:

Assessment context is a property of test items indicating how closely the test items conform to certain features of a real-life, out-of-school problem situation. Operationally, the assessment context of physics test items is related to: 1) The redundancy of essential information for solving the item, 2) The presence of irrelevant information, 3) The number of physics concepts among the response alternatives. An item is high in assessment context, that is, a high context item, if it has redundancy of the essential information, does not have irrelevant data, and has only one physics concept among the response alternatives. A low context item has minimal redundancy of the essential information, at least one example of irrelevant information, and more than one physics concept among response alternatives.

A diagrammatic test item is one having a drawing of a situation in which the concepts and principles of physics may be applied. A few words are sometimes used in diagrammatic items to label or to describe aspects of the diagrams. Diagrammatic items are free response in format. Students answering the items are instructed to look for discrepancies between what they see in the diagram and what they would expect the situation to look like based upon their knowledge of physics. Diagrammatic items are low context items.

The Discrepancy Detection Test (DDT) is a 20-item test composed of diagrammatic items.

The High Context Test (HCT) is a 20-item subtest composed of high context, multiple choice items.

The Low Context Test (LCT) is a 20-item subtest composed of low context, multiple choice items.

A physics concept is defined as an inferred mental process learned by distinguishing positive and negative instances of stimulus objects in a certain class (Gagne, 1970). In this case the domain is the domain of physics. By this definition "weight", "electric charge", "kinetic energy" are concepts. Furthermore "8 newtons weight" and "10 newtons weight" are two examples of the weight concept. However, Coulomb's Law and Archimedes' Principle are not concepts since each contain several concepts within and are not learned by distinguishing positive and negative instances of stimulus objects.

A physics principle is defined as a relational concept or rule (Gagne, 1970). Relational concepts establish a meaningful association of two or more simpler concepts. For example, "momentum is the product of mass and velocity" is a relational concept. Rules are inferred capabilities enabling a predictable class of responses to be made to a given class of stimulus situations. Thus, for Archimedes' Principle the class of stimuli is "Objects placed in fluids" and the class of predictable responses is "Experience buoyant forces equalling weight of displaced fluids"; Archimedes' Principle conforms to the definition of a rule and hence is an example of a physics principle. For the present study physics principles are included within physics concepts.

The redundancy in a physics test item is defined as 1) the presence of information enabling the correct response to be reached by more than one line of reasoning, or 2) the provision of units of measurement or standard abbreviations for concepts as well as naming the concepts, or 3) the naming of principles of physics in addition to specifying component concepts or quantities, or any combination of 1), 2), or 3). Redundancy is explained in detail in Chapter 4.

The definitions related to cognitive style follow:

Breadth of categorization is defined as the range of differing stimuli which are accepted by an individual in the same class or category (Bruner and Tajfel, 1961). Operationally, breadth of categorization is measured by the Category Width Scale (CWS) (Pettigrew, 1958).

Field independence is the extent to which an individual is able to overcome an embedding context or separage an item from its context in a perceptual-conceptual task (Witkin et al., 1962). Operationally, extent of field independence is assessed by the Hidden Figures Test (HFT) (Jackson et al., 1964).

The operational definitions for physics achievement and intellectual ability are presented below:

Physics achievement is assessed by means of the Physics 30 Test (P30), June 1972, as set by the Alberta Department of Education for students studying the Physics 30 course from the textbook by Stollberg and Hill (1968).

Physics achievement-knowledge level is measured by the subset of items of the P30 which are rated at the knowledge level according to Avital and Shettleworth (1968) and which comprise the Avital-Shettleworth Knowledge (ASK) subtest.

Physics achievement-algorithmic thinking level is measured by the subset of items of the P30 which are rated at the algorithmic thinking level according to Avital and Shettleworth (1968) and which comprise the Avital-Shettleworth Algorithmic Thinking (ASAT) subtest.

Quantitative ability is measured by the Cooperative School and College Ability Test (SCAT)-Quantitative, Level 3A.

Verbal ability is measured by the SCAT-Verbal, Level 3A.

Assumptions

Assumptions made in the course of the study are listed below.

1. The assumption has been made that the students responded to the tests as best they were able. This assumption appeared to be warranted since the investigator had full cooperation of the participating teachers and the students gave no overt indications of non-cooperation according to the perceptions of the investigator.
2. It is assumed that the high context items and the low context items are sufficiently different along the assessment context dimension so that the HCT is meaningfully different in assessment context from the LCT and the DDT.

Delimitations

The following delimitations have been accepted for the study:

1. In order to make the study manageable only two cognitive style variables have been investigated.
2. Only cognitive aspects of real-life problem situations have been used in deducing the qualities of assessment context. Although attitudes and values play a part in most real-life situations they are difficult to handle consistently in a deductive process; they are not part of the present investigation.

3. Assessment context has been investigated only for physics items, the reasons being that the investigator's training and experience in this field made possible the design of physics items.

Limitations

The limitations listed below may have affected the generalizability of the results.

1. The various tests were not administered to the various groups of the sample in a counterbalanced plan. The reasons were that the administration of the ability measures and the physics achievement test was not under the control of the investigator. Furthermore, the participating teachers requested that the multiple choice tests, the HCT and LCT, be administered first to give them more time to review the results with their students.
2. The sample was not a random one and is therefore subject to unknown bias. The schools participating included an inner city school, a near-suburban school and a separate school which may have removed to a degree certain obvious socio-economic biases.
3. The administration of the SCAT-V and SCAT-Q took place approximately three years prior to the administration of the other tests. The time lag has undoubtedly reduced the effectiveness of the ability measures as covariates of the study.

4. Only grade 12 physics students participated in the study. This limitation was imposed in part because of the choice of the schools and also because no provincial achievement tests in physics were administered to grade 10 or 11 students.

Summary

The concept of assessment context of test items in physics has been deduced from the consideration of real-life problem situations. Performance in different assessment contexts is hypothesized as having a relationship to the cognitive style dimensions of breadth of categorization and extent of field independence. In order to be made meaningful for educational purposes, the association of assessment context performance and cognitive style has to be described in relation to overall intellectual ability and to achievement in the subject field. To accomplish this task is the purpose of the present study. The main aspects of the design for achieving the purpose of the study have been outlined. Then, the assumptions, delimitations, definitions and limitations of the investigation have been presented. The chapter concludes with an overview of the report.

Overview

The introductory chapter is followed by a review of the literature on field independence, breadth of categorization, relevant aspects of evaluation in physics and of intellectual ability. Chapter 3 presents the rationale behind the assessment context variable and the details of the various testing instruments. The design of the study is also described in Chapter 3. In Chapter 4 the rationale is presented for the hypotheses of the study; and further, the mathematical techniques which will be used in testing the hypotheses are presented. The results and discussion are contained in Chapter 5. Chapter 6, the final chapter, consists of the summary of the study, conclusions, and recommendations.

CHAPTER 2

REVIEW OF THE LITERATURE

In the survey of the literature which follows, research related to the cognitive styles of field independence and category width will be considered first. For each style dimension, background studies will be reviewed after which investigations relating the style to the process of schooling will be analyzed. The section will conclude with an assessment of methods used for measuring the subjects' position on the style dimension. The present chapter also includes a review of studies related to the evaluation of physics achievement.

Field Independence

Wertheimer (1945) in an analysis of problem solving, has stated that the finding of a correct solution requires the separation of problem components from the context of the situation and their recombination to form new relationships. The separation and recombination of components in relation to perceptual functioning has been extensively investigated by Witkin, Hertzman, Machover, Meissner and Wapner (1954) and by Witkin, Dyk, Faterson, Goodenough and Karp (1962). Witkin et al. (1962) pictured field independence as the perceptual component of a broader aspect of intellectual functioning which they referred to as analytical-global.

The analytical person has highly developed ability to overcome embedding contexts and to experience items separately from the field in which they are contained. The global person exemplifies a style of functioning which submits to the overall organization of the field and which experiences items as part of the background (Witkin et al., 1962). Since the pioneering work of Witkin and his associates many investigators have endeavoured to establish and clarify the place of field independence in cognition and the present study examines the role of field independence with respect to performance on physics test items.

Foundations

In the early work of Witkin et al. (1954) three types of tests were employed: the tilting-room-and-chair tests, the Rod and Frame Test (RFT) and the Embedded Figures Test (EFT). Two types of tilting-room-and-chair tests were employed although for both of them the subject is seated in a tiltable chair facing into a room which also could be tilted. The orientation of the chair and the room could be adjusted either by the subject or the experimenter depending upon the test. For the Body Adjustment Test (BAT) the room is permanently tilted and the chair initially also tilted. The subject is instructed to rotate the chair until it is oriented vertically. In the Room Adjustment Test (RAT) the chair in which the subject sits is fixed at an angle and the subject is instructed to adjust the orientation of the room, initially tilted, until it seems to be vertical.

For the RFT the subject is seated in a darkened room.

An illuminated square frame is placed in front of the subject. Contained in the frame is an illuminated rod. With the frame and rod initially tilted at various angles the subject is required to rotate the rod to the vertical. In the RFT as well as the BAT and RAT scoring is based upon the number of degress away from the vertical of the rod or chair or room which the subject has attempted to place vertically. Thus low scores are associated with good ability to carry out the aligning process.

The EFT is a modification by Witkin (1950) of an earlier test developed by Gottschaldt. For each of twenty or more complex figures the subject is required to locate a simple figure which is embedded within the complex pattern. The score on the EFT is based upon the time required to locate the simple figures. Thus low scores are associated with good ability to locate the required figures.

Witkin et al. (1962, chap. 4) reviewed the striking individual consistencies of performances on BAT, RAT, and RFT. They also described the rejection of hypotheses that the consistencies were due to an innate sense of perception of the upright or to body sensitivity, and the eventual retention of the hypothesis that the consistencies were related to the ability to overcome an embedding context.

Further studies reported in Witkin et al. (1962) suggest that the relationships among the BAT, RFT and EFT are stable across an age range from 8 years to adulthood and for many different groups of adults. The RAT ceased to be utilized for assessing field independence when it was shown not to yield results which were consistent across age and subject groups (Witkin, 1967).

Verbal abilities. Attempts to establish the relationship between extent of field independence and other areas of intellectual functioning are presented in Witkin et al. (1962). Positive relationships between field independence and results of the Wechsler Intelligence Scale for Children (WISC) have been found, with field independence more closely related to the performance scale of the WISC than to the verbal scale. Highly field independent children are likely to perform well in the Block Design of the WISC in which the subject rearranges blocks to reproduce a reference design, or in the Object Assembly in which parts must be assembled to make a meaningful picture. In a factor analytic study of the inter-correlations among WISC subtests and perceptual tests for 10-year old children and 12-year old children Witkin et al. (1962) identified a verbal comprehension factor, an attention-concentration factor, and an analytical field approach factor. The verbal subtests figured prominently in the first factor while the perceptual tests were of slight importance. The perceptual tests contributed heavily to the third factor along with Picture Completion, Block Design and Object Assembly. However, comprehension did load to an appreciable degree on the third factor, and the oblique simple structure which was used to interpret the analysis produced correlations of 0.30 and 0.34 between the first and third factors for the 10 year old and 12 year old groups, respectively. The factor analytic study and other studies of Witkin et al. (1962) exemplify attempts to clarify the relationship of field independence and other verbal-mathematical activities.

Employing college undergraduate male and female subjects Bieri, Bradburn and Galinsky (1958) reported that performance on the EFT was positively associated with intellectual variables. Bieri et al. interpreted the finding as being consistent with Thurstone's (1944) suggestion that performances on the Gottschaldt Test, an earlier form of the EFT, were indicative of cognitive functioning beyond immediate perceptual content. Both male and female groups in the study of Bieri et al. (1958) evidenced a significant relationship, $r=-.40$, between the EFT and the Mathematics score on the Scholastic Aptitude Test (SAT) with better SAT performance by the more field independent persons. No relationships of statistical significance were observed between EFT and Verbal SAT scores. The conflicting pattern of results in the years since 1958 has shown that the relationship between extent of field independence and other intellectual abilities is not a simple one.

That verbal abilities are not related to extent of field independence is suggested in studies of male undergraduates (Bloomberg, 1965; Karp, 1963; Wachtel, 1971) and in a study of army males (Fleishman and Dusch, 1971). A positive relationship between verbal ability and extent of field independence is indicated in studies of male undergraduates (Highley, 1970) and in a study of male and female grade 7 and 8 pupils (Stuart, 1967). In the above studies the EFT, the Hidden Figures Test (HFT) which is a modification of the EFT for group administration, or the RFT were used in assessing field independence. Verbal abilities were indicated by scores on the Verbal SAT, Verbal College Aptitude Test (CAT), or a reading grade level score. However, the variables of type of measuring instrument, sex and academic or age level were not utilized systematically so as to provide a basis for an explanation of the conflicting results. The relationship of field independence and verbal ability remains in doubt.

In reviewing the conflicting evidence, Wachtel (1972) has noted the test-construct distinction and has suggested that the shared variance of several related tests may provide a better estimate of the construct of field independence than can any single test; furthermore, since tests of field independence may be related to measures of ability, future studies should control for possible ability effects. Vernon (1972) has also summarized the literature on field independence and, while agreeing with Wachtel (1972) on the desirability of assessing field independence from a small battery of tests, has suggested that the reason for some of the conflicting evidence may lie in the make-up of subject samples. Vernon (1972) notes that some groups may be more homogeneous than others and that correlations of spatial tests and verbal abilities tend to be lower in the more uniform groups.

In the present investigation practical considerations have made necessary the utilization of a single test to estimate field independence. The advice of Wachtel (1972) has been followed in controlling for effects of verbal and mathematical abilities.

Mathematical abilities. Physics test items are administered in the present investigation. Because some physics items require that subjects perform mathematical calculations, the literature concerning field independence and mathematical ability has been reviewed. The picture which emerges is ill-defined, containing contradictory lines.

Statistically significant, positive relationships between extent of field independence and quantitative ability have been reported for college males by Bieri et al. (1958) and Spotts and Mackler (1967) and for college females by Bieri et al. (1958) and Highley (1970). No relationship between field independence and quantitative ability was indicated in studies of male undergraduates (Karp, 1963; Bloomberg, 1965) and male army personnel (Fleishman and Dusek, 1971). Because the evidence is conflicting, the possibility exists of an association between field independence and quantitative ability. Hence the effects of quantitative ability will be controlled in the present study.

Sex differences. Pysh (1970), and Dreyer, Nebelkopk and Dreyer (1964) have concurred with Witkin et al. (1962) in noting no sex differences in field independence for children below 7 years. That males are more field independent than females in the age range from 8 years through adulthood has been suggested by Witkin et al. (1962; 1967) and by Schwartz and Karp (1967). College males are reported to be more field independent than females (DeRussy and Futch, 1971) although Jackson, Messick and Meyers (1964), and Feather (1967) have found no significant sex differences for the same population. Davis (1972; 1973) has also found no significant sex differences in field independence among undergraduate educational psychology students. Jackson et al. (1964) attribute the finding on non-significant sex difference as not being generalizable. They argue that their result could be explained by a positive association between field independence and overall ability, and by selection procedures at the particular college where their subjects were enrolled. The selection procedures could have resulted in the enrollment of females of abilities higher than enrolled males, this leading to a comparison of field independence in average ability males and above average ability females.

In a report of an on-going longitudinal study of cognitive style in relation to academic choice and performance, Witkin (1972) has indicated that even in the high school years a pattern begins to emerge with the more field independent persons predominating in science and mathematics. The trend is more pronounced for females than for males. At the college level Witkin (1972) has found that female mathematics majors were more uniformly high in field independence than male mathematics majors. This finding suggests that whether or not differences are found in the field independence of college males and females may depend upon the subject field from which the students are selected. If females are more field independent than males in a specialty such as mathematics, which is typically chosen by analytical persons, possibly males are less field independent than females in a specialty such as education, which is typically chosen (Witkin, 1972) by less analytical persons.

The subjects of the present study are physics students in their senior high school year. Usually many more males than females are enrolled in physics and Witkin's (1972) findings suggest that the females enrolled in physics are decidedly among the more field independent of their sex. Therefore, sex differences in field independence are not anticipated in the present study.

Recent studies by Davis (1972; 1973) with male and female students from introductory classes in educational psychology have shown consistently the superiority of the highly field independent persons in concept identification tasks. However, the particular point at which field independence functions, whether in general intellectual ability, in memory, in attending to details, or in utilization of feedback remains to be clarified.

Studies by Messick and Fritzky (1963), Messick and Damarin (1964) and Wachtel (1968) suggest that the superior performance of highly field independent persons on concept identification tasks may not be related to memory factors. Messick and Fritzky (1963) and Wachtel (1968) found that field independence was not related to ability to learn and remember names of designs. Messick and Damarin (1964) reported that extent of field independence was significantly negatively correlated with memory for faces.

School-Related Research

The main purpose of the present study is the examination of relationships among the assessment context of physics test items, cognitive style and physics achievement. Students taking grade 12 physics constitute the sample. The research reviewed in this section is pertinent because of the age level of subjects, the nature of variables investigated or because of a combination of these two features.

Grieve and Davis (1971) conducted a study in which grade nine geography achievement was related to cognitive style and to discovery and expository methods of instruction. After studying a unit on the geography of Japan during a three week period under either a discovery or expository method the students were assessed by two different tests, one measuring outcomes at the knowledge level (Bloom, 1956), the other at higher levels. For the entire sample of males and females there was a significant positive relationship between field independence and performance on the higher levels test but no other effects nor interactions were observed. After elimination of the one-third of the sample scoring in the middle range on the HFT, the analysis showed a method by field independence interaction for males on both the knowledge and higher levels tests with the more field independent subjects scoring higher under the expository treatment.

The finding of Grieve and Davis (1971) that the more field independent males showed greater superiority to less field independent males under expository than discovery method concurs with Witkin's (1972) observations. Witkin has reported that the more field independent teachers prefer a lecture method to a discovery method and that superior achievement of pupils occurs when there is a match between pupil and teacher styles.

The procedure of Grieve and Davis (1971) in assessing achievement at the knowledge level separately from the higher levels will be followed in the present study. Such a practice makes necessary fewer distinctions between the various Bloom (1956) categories. Poole (1971) has noted the unreliability of such distinctions. The present investigation focuses on the context of the assessment rather than on the nature of the instructional method; nevertheless, the Grieve and Davis finding that highly field independent subjects exhibited greater superiority on the higher learning test than on the knowledge level test suggests that a similar result should be investigated with respect to physics achievement.

The study of Grieve and Davis (1971) did not assess the role of verbal and mathematical ability in the performance of pupils. Already cited are studies (e.g. Bieri, 1958; Spotts and Mackler, 1967; Stuart, 1967) suggesting that field independence and other intellectual abilities are positively linked. Possibly the superior performance of the more highly field independent subjects on the higher level test is due to superiority in underlying intellectual ability, not field independence. This possibility suggests that the effects of verbal and numerical ability should be controlled in the present study.

Davis and Klausmeier (1970) have reported an investigation in which twelfth grade males who had been assessed on extent of field independence performed a concept learning task. Stimuli having seven variables, each with two values, were presented visually to subjects. The variables were letter (H or L), color (red or blue), size (large or small), number (one letter or two), orientation (upright or tilted), horizontal position (left side or right side), and vertical position (at top or bottom). Subjects were instructed that the concept was to be identified by a particular combination of two dimensions, for example red H's, or tilted L's, or large left-side letters. The complexity of the task was varied by altering the number of the variables used in providing instances and non-instances of the concept. In the simplest test only one non-relevant variable was included among the stimuli in addition to the two necessary variables; then, three non-relevant variables were included, and finally all five non-relevant variables. In responding, subjects chose an hypothesis and were given feedback as to whether or not their hypothesis was correct for that stimulus. Errors to criterion performance were recorded. Two different concept tasks were carried out each at three levels of complexity. Davis and Klausmeier (1970) found that the more field independent high school males did significantly better than the less field independent subjects on one of the tasks but not on the other. When the concept identification problem was difficult the more field independent subjects had a decided advantage, but not when the problem was easier. Altering the complexity of the task, however, did not result in differential effects for more and less field independent subject. The superior performance of the more field independent subjects of Davis and Klausmeier (1970) and Grieve and Davis (1971) is probably not due to superior memory capability since Messick and Fritzky (1964) and Wachtel (1968) both employed undergraduate male subjects and reported no significant relationship between extent of field independence and ability to recall designs seven to ten minutes later. Messick and Damarin (1964) found a significant negative correlation between memory for faces and extent of field independence, the memory test given two hours after learning had occurred.

The way in which concepts of physics are learned by senior high school students and the manner in which achievement measures are made bear a tenuous relationship to the process of concept identification in the laboratory situation. Seldom is the learning of a physics concept one of learning to distinguish instances from non-instances; instead the learning process involves relating a new concept to others previously learned and combining the new concepts with existing concepts to establish more complicated rules. The intellectual processes used in responding to test items to measure achievement in physics appear to resemble more closely those involved in laboratory concept identification than those of concept learning in school. In a concept identification task such as that investigated by Davis and Klausmeier (1970) the subject must examine the stimulus, analyze its details, recall the feedback from earlier stimuli and generate a response hypothesis. In responding to a physics test item the student must examine the item, verbal or diagrammatic, call forth concepts and rules from memory, generate hypotheses which may be checked against data availability or responses provided in order to decide upon the choice of response. With low context test items the processes of item inspection, recalling concepts, and making the tests of hypotheses are likely to be more extensive than for high context items. The non-relevant information and the multiple concept responses of the low context items require a more complex responding process. For difficult items, such as those at the three most complex levels of the Bloom (1956) Taxonomy, the responding procedure is likely to be more complex. In the present study the highly field independent persons can be expected to show superiority to the less field independent persons in responding to physics test items at the higher levels of the Bloom (1956) Taxonomy.

The low context assessment items prepared for the present study may place greater demands on short-term memory than do high context items in that subjects have to remember subsets of data presented in an item while assessing the relevance of other data or the various response alternatives. The evidence of Messick and Fritzky (1964), Messick and Damarin (1964) and Wachtel (1968) suggests that those subjects high in extent of field independence are not likely to have an advantage due to memory capabilities.

Not all high school courses are equally chosen by those who are high in field independence. In a longitudinal study of college students Witkin (1972) reports that the more field independent college students had taken more optional science and mathematics courses in high school than had the less field independent students. DeRussy and Futch (1971) found that college students majoring in mathematics, physics and chemistry were more field independent than those majoring in the liberal arts. Barrett and Thornton (1967) reported that a sample of male engineers and technicians proved to be significantly more field independent than a sample of male undergraduates who were not majoring in physical science. The subjects of the present study are grade 12 students enrolled in physics which is an optional course. Based upon the findings cited above, these subjects can be expected to be more field independent than the average grade 12 student, but the more field independent physics students may not be higher on physics achievement than the less field independent physics students.

Brilhart and Brilhart (1971) found a non-significant, although positive relationship between extent of field independence and class ranking within a group of male engineering students. They speculated, however, that the average score of their sample on the field independence measure was higher than it would have been for all university students. The basis for speculation was that the HFT scores distribution had pronounced negative skewness. A non-significant relationship of field independence and physics achievement may occur for the subjects of the present study, but whether or not this relationship proves significant should not affect differential performance between high and low assessment context items of common physics content.

Measurement

Studies of field independence and school-related implications cited above reveal several examples of inconsistencies among the research reports. The role of the measurement instruments will now be explored insofar as the methods of assessment may account for some of the discrepancies in the findings of field independence with respect to sex differences and to other intellectual abilities.

Instruments

In the initial work on field independence by Witkin et al. (1954) assessments were made with the RFT, RAT, BAT and EFT. The use of the RAT was eventually discontinued since it's assessment of field independence did not converge with those of other instruments, but the evidence for the other tests was shown to be quite convincing in showing that beyond the age of eight years males exhibited a greater extent of field independence than females, and the consistency of measurement was greater for males than for females (Witkin et al., 1962). Since Witkin's (1962) publication few other investigators have reported on the use of the BAT, probably because of the elaborate equipment requirements. The use of the RFT has continued in spite of problems with certain models (Lester, 1968; Vaught, 1969; Stuart and Bronzaft, 1970), and the use of the EFT has increased, particularly since Jackson et al. (1964) reported favorably on a group administered form, the HFT (French et al., 1963). Coincident with this changing pattern of instrument usage the findings on sex differences of adolescents and adults on test performance have been less consistent than prior to 1962.

There is a tendency for sex differences to occur consistently from measurements using the RFT. Besides the studies cited in Witkin et al. (1962) two other reports have been located which have made measurements and reported the means for both sexes on the RFT. In both studies (Vaught, 1965; Sherman, 1967) females were less field independent than males. From measurements using the EFT two studies have been found which reported on mean sex differences. DeRussy and Futch (1971) found statistically significant differences between males and females which Stuart (1967) did not. Sex differences are not usually found for the HFT. Of seven studies (Jackson et al., 1964; Feather, 1967; Willoughby, 1967; Boersma, 1968; Osipow, 1969; Davis, 1972 and 1973) six report no significant difference between means for males and females. Osipow (1969) employed a different version of the HFT, and the resulting means for 328 females and 37 males were 87 and 93, respectively, with standard deviations in the range of 15 to 20. While statistical tests on such widely different sample sizes are not appropriate, the differences of the means in relation to standard deviations is not large. All of the studies recounted in this paragraph involved subjects older than 11 years, in most cases college undergraduates.

The evidence indicates that recent measurements on adolescent and adult subjects by means of the HFT do not yield sex differences on mean scores. Either the HFT measures a different capability than the RFT or cultural changes have reduced sex differences in field independence. Both possibilities are discussed below.

Elliot (1961) has stated that performance on the RFT tends to be independent of verbal and quantitative ability measures whereas field independence, as assessed by the EFT, tends to have significant relationships with these abilities, especially quantitative abilities. A study of male undergraduates (Elliot, 1961) bore out the predicted relationship. Since the HFT is adapted from the EFT (Jackson et al., 1964) the relationship predicted by Elliot (1961) could be expected to apply to the HFT. Elliot's (1961) study clearly implies that whatever is measured by the RFT is not the same as that measured by the HFT. Although subsequent research has not consistently replicated Elliot's (1961) findings the conclusion that the RFT and HFT assess somewhat different constructs appears warranted.

Thornton and Barrett (1967) have pointed out that correlations between scores of the RFT and the EFT have not as a rule been as consistently high for females as for males and the claims that, for females, the two tests equally well assess field independence may be unwarranted. However, Denmark et al. (1971) have noted that Witkin's (1954) correlation of 0.76 for college males is much higher than the correlation of 0.43 which they found for a similar sample. Elliot (1961) has reported an EFT-RFT correlation of 0.42 for college males. Making the picture more confusing, Dubois and Cohen (1971) have found a correlation between the EFT and the RFT of 0.56 for a sample of college females. All of the correlation cited above are statistically significant, $p < .01$. The finding of Dubois and Cohen (1971) taken together with those of Elliot (1961) and Denmark et al. (1967) suggest that females may not be less consistent than males in performance on the two tests. While the correlational studies cited in the present paragraph cast doubt on the stability of sex differences in EFT-RFT performance they do not contradict the hypothesis that the EFT, and presumably the HFT derived from it, measures something different from the RFT.

Dubois and Cohen (1970) have calculated the correlations of the EFT, RFT, and various ability measures for a sample of female undergraduates and have found that the EFT correlates just as highly with various ability measures as with the RFT. In addition the RFT-ability correlations are also significant, $p < .01$, although not as high as the EFT-ability relationships. Highley (1970) has also found significant correlations between RFT, mathematical and verbal abilities for a sample of college females. Dubois and Cohen (1971) have suggested that various field independence measures may not be assessing the same contruct. Their suggestion reinforces that of Elliot (1961) who noted that the EFT is like an aptitude test in that it is timed, and also that its items resemble those of aptitude tests. The RFT, according to Elliot (1961), provides a minimum of cues which might arouse concern because of evaluation, and the RFT is not timed. Thus the RFT is less likely to be multifactorial than is the EFT. The HFT, because of its group administration procedure, may be even more like the group administered aptitude tests commonly written by high school students than is the EFT. Hence, correlations of HFT scores and verbal and numerical ability measures could be expected larger than correlations of the EFT and ability measures.

The foregoing analysis suggests that changing patterns of sex differences on various tests used to assess field independence may be due to inconsistencies in what the various instruments measure. Another possibility is that cultural changes have occurred tending to increase the extent of field independence in females. This, coupled with increasing use of the group-administered HFT, could produce the pattern which was referred to above. This pattern reveals sex differences in field independence as assessed by the HFT since 1964. Wachtel (1972) has pointed to a dichotomy in the interpretation of field independence.

According to Wachtel (1972) adaptive ability needs to be distinguished from adaptive capacity. Thus, a person may have a certain capacity of field independence but that capacity may or may not have developed as the person matured. Hence, an assessment of field independence of persons at the high school or college level may for one person measure the result of a high development of a rather limited capacity, for another the result of a poor development of a larger capacity, for another, medium development of an average capacity, etc.

Cultural influences which tend to develop field independence more for one sex than the other might contribute to observed sex differences. Witkin et al. (1962) have shown that the mother-child relationship can influence the development of field independence in children and have argued that both genetic and cultural factors are involved. Vaught (1965) and Sherman (1967) have claimed that the cultural factor dominates, and in a direction which results in greater field independence for males than for females. If Sherman (1967) and Vaught (1965) are correct, and if a cultural change has occurred resulting in smaller sex differences for college students assessed on field independence since 1964 as compared with before, the cultural change explanation for the observed pattern of sex differences in field independence measures could be accepted. To the investigator, the cultural change explanation is less persuasive than that involving the possibility of instrumental inconsistencies.

The trend toward no observable sex differences in field independence of college students with the HFT may be more apparent than real because of sampling factors. In a longitudinal study of students throughout the high school and college years Witkin (1972) has found that college students assessed as high in field independence had taken more high school courses in mathematics and science than those low in field independence. Witkin (1972) has also noted that females who are college majors in mathematics are more consistently high in field independence than are males in such programs. Furthermore, Witkin has documented that students in education are decidedly among the less field independent and that education is preferred by more women than men. Conceivably a selection process has been at work in which a large fraction of the more field independent males and a small fraction of the more field independent females are enrolled in specialties such as mathematics and science, while in a subject area preferred by less field independent persons, such as education, fewer males than females are enrolled with the males constituting a less diverse group than the females. This analysis could account for the less populous but less diverse group of highly field independent females than males in majors such as mathematics (Witkin, 1972) and for the unimportant sex differences in

HFT scores of the introductory educational psychology subjects of Boersma (1968) and Davis (1972; 1973).

The following comments are presented in summary of the section thus far. The RFT, BAT, EFT were the instruments most commonly used prior to 1964. Few investigators other than Witkin et al. (1954; 1962) have used the BAT because of elaborate equipment requirements. Although apparatus problems have been encountered with the RFT (Lester, 1968) this instrument continues to be used in individual testing while a group-administered form of the EFT, the HFT (French et al. 1963) has been increasingly used since 1964 (Jackson et al., 1964). Sex differences in college subjects are usually found in RFT scores but not in HFT scores. Furthermore, the HFT variable usually correlates more highly with verbal and numerical ability measures than does the RFT. Suggestions that the RFT and HFT assess different constructs and that the HFT may be multifactorial in nature appear to be warranted. The results on mean sex differences or lack thereof for the RFT and HFT are interpretable in terms of the multifactorial nature of the HFT and sampling considerations. One other aspect of instrumentation remains to be discussed in this section, that of practice effects on HFT performance.

Jackson et al. (1964) have discussed the phenomenon of practice effects for EFT items and have noted that practice has an important effect, so much so that without order of presentation information, item difficulty indices are meaningless. Boersma (1968) investigated practice effects on the HFT, a form of the EFT which is nearly identical with that designated by Jackson et al. (1964) as the EFT Form III. The HFT consists of two separately timed parts each containing 16 items. Ten minutes are allowed for each part. Boersma (1968) administered the HFT to 105 male and female first year education students on two occasions separated by a 10-week interval. Table 1 presents the result of various investigations of practice effect, including Boersma's.

Boersma found an increase in the mean score from 4.58 to 6.06 on the two parts of the test during the first administration, and scores of 8.31 and 7.65 were recorded for parts 1 and 2 of the second administration. Mean total scores of 10.64 and 15.96 were recorded on the first and second administrations, respectively. Boersma (1968) did not correct the scores for guessing. Fleishman and Dusek (1971) corroborated the results of Boersma in their study of adult male army personnel. The results of Fleishman and Dusek's study showed an increase in mean total HFT scores corrected for guessing which went from 10.50 in the first trial to 23.45 on the fifth trial and declined to 21.20 on the sixth trial. The trials were spaced at one-day intervals. Brilhart and Brilhart (1971) administered the HFT to a sample of engineering students at the end of their first, third and ninth terms in college, with gaps of approximately one-half and two and one-half years between administrations.

Table 1

Summary of the Evidence for Practice Effects in the Mean Scores on the HFT

		Administration					
Investigators	N	1	2	3	4	5	6
Boersma (1968)	105	10.64	15.96				
Fleishman and Dusek (1971)[a]	10	10.50	13.10	15.08	18.58	23.45	21.20
Brilhart and Brilhart (1971)	148	12.00	12.00	11.83[b]			

[a] Scores corrected for guessing.
[b] N = 105 for this administration.

Their results, uncorrected for guessing, showed no practice effect.

The most obvious explanation for the inconsistency of the results between Brilhart and Brilhart (1971) and the other investigators lies in the longer time which elapsed between the administrations by Brilhart and Brilhart (1971) as compared with those of Boersma (1968) and Fleishman and Dusek (1971).

Jackson et al. (1964) have expressed the idea that the characteristics measured by an embedded-figures test after considerable practice might be different from those initially assessed. A few investigators have reported means scores on the HFT for college undergraduates which are very much higher than those obtained in the initial administrations of Boersma (1968), and Fleishman and Dusek (1971). Davis (1972), for example, reported means of approximately 24 after the scores had been corrected for guessing. In the absence of information on any alteration in the usual administration procedures for the HFT one concludes that the subjects may have taken the test previously, possibly several times in order to achieve scores similar to those of Fleishman and Dusek's (1971) subjects on their fifth trial.

Summary

The present investigation examines relationships between criterion tests in physics which vary in assessment context and the independent variables of cognitive style and physics achievement. Because of its low cost, ease and quickness of administration, and capability of group administration the HFT has been chosen in preference to the EFT or the RFT.

Research (Bieri et al., 1958; Highley, 1970; Stuart, 1967; Spotts and Mackler, 1967; Lezotte, 1969) has suggested that extent of field independence is positively related to verbal and numerical abilities. Positive relationships are particularly likely to be found when field independence is assessed with the EFT or one of its modifications. Vernon (1972) points out that the EFT is likely multifactorial in nature. Since the HFT is a modification of the EFT Vernon's conclusion probably applies to the HFT as well as to the EFT. The possibility of overlap in measurement of the HFT and verbal and numerical ability measures is taken into account in the present study in that effects of verbal and numerical ability will be controlled when the HFT is used for predicting criterion performance.

Although significant sex differences have been reported in extent of field independence (Bieri et al., 1958; DeRussy and Futch, 1971; Witkin et al., 1962) important sex differences have not been reported in mean scores of the HFT (Jackson et al., 1964; Boersma, 1968; Davis, 1972; Davis, 1973). Since the HFT is used to measure field independence, the effects of field independence are not assessed separately by sex.

Assessment context of test items is varied in the present investigation. Low context items, as contrasted with high context items, contain less redundancy of relevant information, no non-relevant information and greater diversity of likely responses. It is hypothesized that low context items may require response processes similar to those for answering items assessing the higher categories of the Bloom (1956) Taxonomy, whereas the high context response process may be more like those required for answering items which assess the knowledge level of the Bloom Taxonomy. Scores on physics achievement test items are anticipated to be more closely related to performance on high context items than with low, while scores for higher category items will likely be more closely associated with low context performance than with high.

While the study of Grieve and Davis (1971) centred on the relationship of field independence and extent of learning under two different treatments, the present study investigates the relationship of field independence and the assessment context of test items. Since the Grieve and Davis (1971) investigation has shown that the Bloom (1956) category of the test item is a factor which differentially affects the performance of more and less field independent persons, the present study has been designed to test hypotheses on the relationship of the Bloom (1956) categories and field independence.

The experiment of Davis and Klausmeier (1970) showed that the more field independent subjects performed better on the more difficult concept identification task, which suggests that the more field independent subjects may achieve better on the low context items than on the high context items, insofar as the low context items are more difficult. Witkin (1972) has suggested that students who enrol in optional science and mathematics courses in high school are usually high in field independence. The subjects of the present investigation (physics students in their final year of high school) are probably more field independent than the population of all senior students.

Breadth of Categorization

Breadth of categorization, or category width, refers to the range of different stimuli which a subject classifies as the same. A person who rates 6 of 10 audible tones of different frequency as having the same pitch is said to demonstrate greater breadth of categorization than a person who rates only two of the 10 tones as the same in pitch. Selected studies which have contributed to the developing and refining of knowledge of category width are reviewed below. Included with the foundational studies are instrumental considerations and a discussion of possible sex differences in breadth of categorization. Studies which seem to have implications for schooling are detailed in a separate section. Throughout the review, possible relationships with the present investigation will be explored.

Foundations

Bruner, Goodnow and Austin (1956), who note the ubiquity of the categorization process in all cognitive activity, have distinguished between responses which result in the formation of categories and those which classify stimuli as belonging to one category or another. The relationship of these two types of categorizing behaviour has been an underlying issue in much of the work on breadth of categorization. In the present study breadth of categorization is investigated in relation to the performance of high school physics students on physics test items which vary in assessment context. The high context items, all of which are verbal, multiple-choice items, are characterized by having response alternatives involving only one physics concept. For example, the four response alternatives "10 kg, 15 kg, 20 kg, or 25 kg" are all viewed as examples of the concept of mass.

In contrast, the low context verbal items are constructed so that several different concepts must be considered in making a response. For example, the response alternatives "10 N, 15 J, 20 kg, or 25 sec" encompass concepts of force, energy, mass and time. In neither type of test item is the subject called upon to form the response categories; however, the process of deciding among different categories or concepts in low context items may call upon different categorizing strategies from those required in making responses in high context items. The Pettigrew Category Width Scale (CWS) is used for estimating the breadth of categorization of subjects in the present investigation.

Pettigrew Category Width Scale. Pettigrew (1958) developed a 20-item, pencil and paper measure of category width. Each item presents the subject with the statement of an average value of some quantity. The subject then makes two selections from among two sets of four alternatives. From the first set the subject chooses an estimate of the maximum-sized incidence of the quantity, and from the second set an estimate of the minimum. A sample item is presented below.

> When all of the world's written languages are considered, linguists tell us that the average number of verbs per language must be somewhere around 15,000. What do you think:
>
> a. is the largest number of verbs in any single language...
>
> 1. 21,000..... ( ) 3. 50,000..... ( )
> 2. 18,000..... ( ) 4. 30,000..... ( )
>
> b. is the smallest number of verbs in any single language...
>
> 1. 1,000..... ( ) 3. 5,000..... ( )
> 2. 13,000..... ( ) 4. 10,000..... ( )

The sizes of the estimates in each direction were derived from statistics accumulated from a free-response form of the test which was administered to 750 college students. The 10th, 35th, 65th and 90th percentile choices larger than the given average value were selected for the four estimates of the maximum. The same set of percentile choices smaller than the average value were used for the four minimum alternatives presented for each item in the final form of the test.

Scoring of the CWS is accomplished by summing the weights of responses to largest estimate and to the smallest estimate for all items. Weights of 0, 1, 2, or 3 are applied to the four maximum and four minimum estimates according to distance from the given average with the response value which is closest to the average in each case receiving the weight of 0. The maximum attainable score in 20 items is 120.

The criterion validity of the CWS was determined by having 26 undergraduate subjects undergo five different assessments of category widths based upon perceptual tests. The tests included the estimation of line lengths which fell within a given size category, and the comparison of weight extremes of ostrich eggs with fixed weights. The subjects then were administered the CWS. There was a statistically significant, $\underline{p} < .02$, degree of consistency across the five perceptual tests. The CWS scores of the nine subjects having broadest category width on the perceptual tests were compared with the CWS scores of the nine subjects having the narrowest category width on the perceptual tests. The mean scores of the two groups were compared for each item of the CWS and the mean differences were compared using one-tailed $\underline{t}$-tests. Chance probabilities of the observed differences were less than five percent for eight of the items, with the largest probability value for any item on the test being 40 percent.

Gardner and Shoen (1962) have pointed out that each item of the CWS and each of the tasks used to validate the CWS assess the limits of one conceptual dimension, and that tasks may or may not be related to other types of categorizing behaviours in which a collection of heterogeneous items are sorted into a complex of more or less related dimensions. The two types of categorizing tasks are discussed in the section below.

Sorting and categorizing. In the sorting task of Gardner (1953) a subject is required to sort a collection of 73 objects into groups or collections which to the subject seem appropriate. Subjects vary widely in the number of groups utilized but the individual stability in groups used seems high. Over a three-year period a correlation of 0.75 was obtained in one investigation of the number of groups utilized (Gardner and Long, 1961). It has been argued that a broad category width in classifying stimuli along one dimension is equivalent to the tendency to sort objects into a small number of groups (Gardner et al., 1959, chap. 5). Presumably the subject who employs a small number of groups in the object sort must be placing together in the same class objects of a wide range of characteristics. The subject has, therefore, a broad category width or equivalence range. However, when responses to the Object Sorting Test were analyzed (Gardner et al., 1960, chap. 6) according to level of abstraction employed, apparently the defining of groups in which to place objects involved different aspects of conceptualization from that required for deciding upon the inclusiveness of a given grouping. To delineate the processes involved, Gardner and Shoen (1962) conducted a series of studies in which both the object-sorting and width-discriminating tasks were employed.

The results of the first study (Gardner and Shoen, 1962) showed that scores of the 70 adult, female subjects on the CWS were not significantly related to the number of groups formed in the object sorting task although the correlation was, as anticipated, negative. In a subsequent study two other pencil-and-paper range-width tests were included (Gardner and Shoen, 1962). Forty of the 70 subjects of the first study formed the sample for the second.

In the second study the number of categories in the object sorting task, one factor of the CWS, and the two added range tests of Fillenbaum (1959) were significantly related, $\underline{r} = .30$. Pettigrew (1958) performed a centroid analysis of the inter-item correlations of the CWS and identified two factors, one attaining significance with object sorting groups and other scores in the Gardner and Shoen (1962) second study. In view of conflicting evidence on the relationship of object sorts and CWS between the two studies of Gardner and Shoen (1962), the relationship between the two types of categorization requirements seems unclear.

A study by Sloane, Gorlow and Jackson (1963) sheds light on various types of categorization tasks. Sloane et al. (1963) employed female undergraduates as subjects and found statistically significant positive relationships between the number of groups formed in a wide variety of sorting encompassing objects, pictures of faces, descriptions of people, and words. Tests designed to assess widths of categories in the realm of size and of shape were, however, not significantly related to number of groups in the object sorting task. Regrettably, from the point of view of relevance to the present investigation, Sloane et al. (1963) did not include the CWS among the tests of their study. Nevertheless, the Sloane et al. investigation confirms the results of the first study of Gardner and Shoen (1962) in suggesting that somewhat different cognitive processes are involved in assessing the width of a given realm, as called for in the items of the CWS, from those required in forming realms as required by the free-sort tasks.

Of the two types of assessment context tests employed in the present study, the high context test appears more closely related to the range width type of assessment than to the sorting type of assessment. High context test items involve only one physics concept among the various response alternatives while the low context verbal test has at least two concepts among response alternatives and the low context diagrammatic test presents no ready-made alternatives. For high context items, response judgements are made within a given realm while the low context items require judgements among two or more realms. In neither type of test, high or low in assessment context, should breadths of categorization be a factor in the responses to items of physics content for the student who is correct with confidence on each item. For him, judgements about which response to make within which realm are not determined by estimating but by a simple comparison between his hypothesized or calculated response and the alternatives for selection. However, since not all students are perfectly confident, and since the assessment context items have been constructed with difficulty levels approximately equal to 0.5, uncertainties are sure to be involved, and the student may evaluate response alternatives with respect to the risks of being right or wrong.

Risk-taking. The effects of risks on behaviour in categorization has been extensively investigated by Kogan and Wallach (1964). In the decision to include or not to include a stimulus within a response category two types of risks are involved. To risk a type I error is to include an inappropriate stimulus with the category; to risk a type II error is to reject from inclusion in a category a possible exemplar. In their study employing 114 male and 103 female undergraduates, Kogan and Wallach (1964) found that for males no relationship existed between category width as assessed by Pettigrew's (1958) factor I of the CWS and various decision making variables. For females however, the risks involved did relate significantly to categorization behaviour. Females who were broadest in category width tended to be those who were willing to pay for more evidence in a number judgement game before committing themselves to a decision which could result in monetary loss or gain for themselves. They tended also to be the females who paid for clues which reduced their chances of losing in a game of object identification. These relationships were especially strong for females who were less confident in their judgements. Kogan and Wallach (1964) interpreted their results as being contrary to earlier suggestions that females were more conservative than males (Gardner and Shoen, 1962) because of narrower average category widths. Wallach and Caron (1959) have speculated that a cultural requirement that females be less forthright than men might be the basis for the narrow category width of females than males. However, in light of Kogan and Wallach's research, another culturally based factor may be the reason for the observed differences, the proclivity for males to acquire greater experience than females in quantitative thinking (Pettigrew, 1958; Gardner and Shoen, 1962). This factor may be particularly apparent with the CWS because of the quantitative nature of its items.

Sex differences. While significant sex differences in breadth of categorization have been reported for college undergraduates on the CWS (Pettigrew, 1958; Rosen, 1961; Bieri, 1969) and for senior-year high school students on the CWS (Field and Cropley, 1970), other investigators have reported contrary results. Non-significant, $p < .05$, relationships for sex differences in CWS means have been found with undergraduate subjects (Feather, 1967a, 1967b; Eagly, 1969) and with pupils aged seven to 11 (Penk, 1969). Also Tajfel, Richardson and Everstine (19[illegible]4) reported no sex differences in performance of college undergraduates on a judgmental task to assess breadth of categorization.

The major implication for the present study seems to follow from Kogan and Wallach's (1964) finding that female undergraduate subjects of broad category width tend to follow a conservative strategy where decisions must be taken without a high degree of confidence. A school testing situation involving difficult items may be an example of such a situation. The implications of a conservative strategy in responding to difficult test items are presented below in the section on research related to schooling.

School-Related Research

Field and Cropley (1970) investigated possible relationships between achievement on a general science test and breadth of categorization as assessed by the CWS. The subjects of the study were male and female students in their senior year of high school. While no statistically significant relationships emerged a tendency was noted for females who were high in science achievement to be low in breadth of categorization. Field and Cropley's results suggest that in the present investigation breadth of categorization and achievement in physics are not likely to be related. Since more males than females are enrolled in senior physics the sample for the study will undoubtedly contain more males than females and whatever tendency may exist for females of narrow category width to score higher in science achievement is unlikely to emerge in the sample. The Field and Cropley (1970) study does not, however, provide an indication of relationships among breadth of categorization and performances on the high and low assessment context tests.

Mathematical and other abilities. In developmental work on the CWS, Pettigrew (1958) performed a centroid analysis of the item inter-correlations. The four factors which emerged were rotated orthogonally. Pettigrew had anticipated that two meaningful factors would appear, one a factor characterized by heavy loadings for items related to the every day experience of subjects, the other with heavy loadings for items far removed from experience. Of the four factors which emerged, two were interpreted as error since their heaviest loadings came from the items with lowest criterion validity. Pettigrew was not able to interpret the two apparently meaningful factors in terms of the everyday-experience and remote-experience hypotheses. Correlations between the American Council on Education test of quantitative ability and factor I and factor II yielded values of 0.33 and 0.06 respectively for the 270 undergraduate subjects.

The relationship of quantitative ability and the CWS has been further explored by Messick and Kogan (1965). Three 15-item tests of quantitative ability were developed. Two of the tests had multiple-choice formats, the other a free-response format. The item stems were similar for the three tests but one of the multiple-choice tests had choices which were widely spaced in that the largest value was several times greater than the smallest, while the other multiple choice test had more narrowly spaced responses. Messick and Kogan (1965) did not report the criterion for distinguishing the two types of spacings but an example of each was provided. For the narrowly spaced example the largest alternative was 1.4 times as large as the smallest while the largest alternative of the widely spaced example was 5.5 times the smallest. The mean difference between the five response "steps" of the narrowly spaced item example was 0.5 while for the largest it was about 200. In the Messick and Kogan (1965) study 40 male undergraduate students in psychology took the three arithmetic tests and the CWS. The correlations between the CWS factor I and factor II with the test having widely spaced alternatives were 0.15 and 0.36, respectively. Only the second coefficient is statistically significant, $p < .05$. None of the other correlations between the two factors of the CWS and the arithmetic tests was statistically significant. Messick and Kogan's (1965) finding contradicts that of Pettigrew (1958).

on the relationship of quantitative ability and the two factors of the CWS. Messick and Kogan attribute the significant relationship between factor II of the CWS and the test with widely spaced alternatives as being due to a common approximation strategy which subjects might be using and not related to computational ability. Implications of an approximation strategy will be discussed below in the section concerning breadth of categorization and originality.

Factor II of the CWS was among the variables studied by Kogan and Wallach (1964). Near zero correlations were obtained between factor II and SAT-Mathematics for 114 college males and also for 103 college females. The studies of breadth of categorization and mathematical ability have not yet revealed a reliable relationship between the variables.

Among the variables of the present investigation are verbal ability and extent of field independence. Studies suggest that performance on the CWS is unrelated to verbal ability as assessed by the SAT (Kogan and Wallach, 1964) to verbal IQ as assessed by the American College Test (Jackson and Pedersen, 1965) and to field independence as measured by the HFT (Messick and Damarin, 1964). The subjects in each of the investigations just cited were college undergraduate males and females.

Originality. Originality and breadth of categorization have been studied by Anderson and Cropley (1966) and Field and Cropley (1970). The subjects for the Anderson and Cropley study were 320 grade seven pupils. A total of 22 measures of originality, ability, and personality functioning were administered. A principal components analysis was performed followed by graphical rotation of the factors to an oblique relationship. Based upon loadings of variables for the originality factor, with major contributions coming from the Guilford-Torrançe tests, the data for the 31 most original and the 31 least original subjects were selected for further study. In multiple regression analysis of the contribution to prediction of originality the CWS was a highly significant predictor alone, $p < .005$, and was significant, $p < .02$, in improving prediction beyond that which was possible from the sex variable. Sex did not, however, significantly add to the prediction of originality beyond that which was possible from the CWS variable.

In studying the relationship of cognitive style and science achievement with senior-year, high school students, Field and Cropley (1970) included among the measures administered the CWS and the Uses of Objects Test (Hudson, 1966). Scored for originality, the Uses of Objects Test was significantly related to the CWS, $p < .05$, for the 104 male subjects, but not for the 74 females. While both the Field and Cropley (1970) and Anderson and Cropley (1970) studies have found a positive relationship for breadth of categorization and originality the non-emergence of sex as an important factor in the Anderson and Cropley study may have been because they were working with subjects selected from the extremes in originality. Anderson and Cropley suggest that the relationship of breadth of categorization and originality may occur through a common risk-taking tendency.

Inspection of the items on the CWS suggests an alternative explanation. For almost every item on the CWS one can imagine a situation in which the extreme choice is the correct choice. For example, one item states that the average flight speed of birds is 17 miles per hour and the maximum speed choice for the fastest bird is 105 miles per hour. Surely at some time or other a fast flying, diving bird with a tail wind has achieved such a speed. The smallest alternative choice for flying speed is 2 miles per hour. Here again the combining together of several factors in an unusual way, for example a humming bird with a head-wind on a cold day, and the smallest alternative appears to be the one to choose. Insofar as the realization of unusual factors and the combining of factors in unusual ways constitutes originality, an explanation of the positive relationship of originality and breadth of categorization is possible.

The finding that subjects of broad category width show superiority on arithmetical items with widely spaced alternatives was cited in the preceding section. Originality may play a role in the relationship. The more original thinker having a wider number of alternatives by which to approach the problem may "see" methods for quickly eliminating some alternatives without calculation, methods not available to the less original subject. For the person who has arrived at the response by elimination, or who has reduced the number of possible responses by elimination, the eventual calculation which is made is one of confirmation and the calculation is not the sole resource available for deciding on the best response.

The items of the tests low in assessment context, containing as they do non-relevant information in the item stems or diagrams and presenting or suggesting diverse response alternatives, may prove to be the items in which the more original thinker may be superior. Therefore a positive relationship is indicated between the low context tests and the CWS on which the more original subjects score more highly.

Summary

Studies have shown (e.g. Gardner and Shoen, 1962) that category width as determined by the number of different items placed together in item-sorting tasks is not closely related to category width as measured by the range of stimuli differing along one dimension which are determined as being the same as a given standard. The CWS consists of items which seem to reflect the latter interpretation of category width.

Kogan and Wallach (1964) have shown, particularly for females making judgements in situations where they are not confident, that the broad categorizers tend to adopt a conservative strategy. Faced with a difficult test item in a school situation, a subject would probably be acting conservatively by guessing at an answer rather than omitting the item when there is no penalty for guessing. For the tests used in the present study no guessing penalty is advertized or imposed. However, the slight advantage this might give broad category width persons on difficult items is not likely to be large since a minority of the physics students are females and since the average difficulty levels of the criterion tests are approximately equal.

Studies cited in the preceding sections suggest that physics achievement, the HFT, SCAT-V and SCAT-Q all of which are predictor variables of the present investigation are unlikely to be related to the CWS variable.

The CWS variable is likely to be related to performances on the assessment context tests. The high context test, which resembles the physics achievement test, will probably not have a significant correlation with the CWS, but the low context tests may be expected to show a significant positive correlation with the CWS. The ability to make distinctions among data and response alternatives of low context tests, an ability once thought to be related to narrow category width, should not alter the positive relationship of the CWS and performance on the low context tests. For example, Bieri (1969) has shown that college subjects low in scores on the CWS are less able to make accurate discriminations of line lengths than their peers who show high scores on the CWS.

Evaluating Physics Achievement

The assessment context variable of the present study has been defined in Chapter 1 based upon real-life application situations. The purpose of the present section is to review the means of evaluating application and thereby to clarify the place of assessment context in the framework for evaluating the application objective.

The methods of evaluating the application of knowledge of physics have depended upon the definition of what constitutes application. For the teachers of the Eight-Year Study (Smith, Tyler et al., 1942) application related to the existence of technical devices or situations in the homes or surroundings of the students; for Bloom (1956) application includes not only the workaday situations but also extensions of that which the students had done in the learning situation of the classroom; and for Hurd (1973), application is a more generalized concept related to the availability of multiple approaches and methods for problem-solving. While the application conception by Hurd (1973) seems the closest to that adopted in the present study of assessment context, the studies cited above, and others too, are reviewed below.

Applications derived from the physical reality are discussed separately from applications based upon the intellectual requirements of the assessment task.

Applications Taken from Concrete Situations

Appraising the progress of secondary school students was an important part of the Eight-Year Study (Smith, Tyler et al., 1942). The science teachers participating in the study ranked high in importance the objective of application of principles of science (Smith, Tyler et al., 1942, pp. 77-111). Test items to assess the application objective were required to meet three criteria: 1) They had to be new to the students, 2) They had to occur frequently in actual life, and 3) They had to be explainable by principles actually studied by the students. Among the examples of application problems presented by Smith, Tyler et al. (1942) are problems about the skidding of an automobile on wet pavement, the buoyancy of an oil drum, the stopping distance of cars at various speeds, and the changing air pressure in tires with air temperature fluctuations.

While the examples meet the three criteria they seem inadequate as means of assessing whether or not the students had learned to apply knowledge to solving problems as they arise in daily living. The reason for the apparent inadequacy is that the statements of the problems are stripped of most of the additional data which inevitably accompany real-life situations. The problem on the skidding automobile does not mention the time of day, the curvature, width or slope of the road, or the speed of the car, all of which are potentially relevant to the question of how to stop the skid. The presence of a surfeit of data from which a selection must be made, a characteristic of the low context test items of the present study, is a part of almost every real-life situation.

Bloom (1956) has noted the difficulty of simplifying application situations which are new to students while preserving reality. For a problem in the cognitive domain some simplification of most real-life situations is certainly necessary. Smith, Tyler et al. (1942, pp. 92-94) and Nedelsky (1965, pp. 31-32) have explained that attitudes and values play a part in arriving at the solution to many real-life problems. The low context test items are simplified from the real-life environment to the extent that the solution should not depend on the values of the respondent. They are not simplified however to the extent that only the data needed for arriving at the solution are presented. For all low context items at least one non-relevant datum is present in the item stem of the verbal items, or in the drawing of the diagrammatic items.

Generalized Basis for Deriving Assessment Items

Hurd (1973) has described the application of knowledge as recognizing the principles of science in relation to daily events and the taking of actions based upon the knowledge. As an example of application Hurd (1973) offers the making of a decision on the need for a nuclear power generating station. The general feature of such problems is the lack of a stable set of conditions. For example, in one set of circumstances the need for a nuclear power station might have to be established mainly on the basis of trends in power consumption. In another situation, if the plant were proposed in a large city, the decision might be based predominantly on safety features for containing radioactivity. Because of the lack of a stable set of circumstances for the problem the student requires training in science emphasizing multiple approaches and widely-varying methods. Hurd's (1973) interpretation of the application of knowledge in terms of varying sets of data provides a basis of assessment which seems related to the low context items of the present study. Not only do the low context items possess non-relevant data in the item stems, but they have response alternatives suggesting competing principles. Just as the decision on the nuclear generator may have to be made on either the demand for power or the requirements of safety, so a test item on a moving object may contain response alternatives calling for a decision based on momentum, or on kinetic energy. With the problem of the nuclear generating station the basis for reaching a decision may depend on a value judgement. Does the basis for solving the low context item then also depend upon a value judgement? The answer is no because of a logical characteristic of the items. For only one response alternative is there consistency with the available data. That is, while data in the item stem may suggest the possibility of calculating either momentum or kinetic energy the data may be complete enough to calculate only one, or, other specifications in the item stem may indicate that only one is acceptable even though both could be calculated. Hence, low context items have a semblance of the many possible variables Hurd (1973) claims must be handled without having the variables associated with the affective domain which can lead to response ambiguity.

Given the claim that the low context items possess elements for the assessment of the application objectives, does that mean that low context items are exclusively application items? The evidence summarized below suggests that the answer is no, but with qualifications.

Bloom's (1956) claim that knowledge, understanding, application, analysis, synthesis and evaluation form a hierarchy has been studied extensively. Kropp et al. (1966) noted that validation studies of the hierarchy would be difficult because of problems in finding response measures which were unequivocally suited to a particular taxonomic level, in establishing response conditions which preclude subject differences in initial knowledge and which do not add knowledge in the process of item responding, and in finding statistics suitable to assess the hierarchical structure. Smith (1970) attempted to solve the first of the three problems above by providing subjects with a communication on a topic, electrical circuitry, with which they had no previous experience and from which test items were derived. Response conditions were chosen so that the parts of the communication related to the simpler levels of the Taxonomy did not include statements related to the more complex levels of the Taxonomy. Smith (1970) employed Hierarchical Syndrome Analysis (McQuitty, 1966) designed to classify taxonomic levels based upon the statistical distance between them. Smith (1970) found qualified support for the Taxonomy with the knowledge, comprehension and application tending to form a cluster and the analysis, synthesis and evaluation levels tending to form a separate inter-related cluster. There were variations within the two-cluster pattern for each of the three subject samples, university undergraduates, ninth-grade students and seventh-grade students. In another study Smith (1971) found that for eleventh-grade students the assumption of no previous knowledge of non-curricular content of economics and social studies may not be valid. The finding of Smith (1971) suggests that the subjects of Smith (1970) may not have been operating at the predicted taxonomic levels, particularly the older, undergraduate subjects.

Stoker and Kropp (1971), assuming no previous knowledge for subjects and providing communications in the form of reading passages, analyzed responses to taxonomic tests by

Guttman-Lingoes (1954) smallest space analysis. If a hierarchy exists of the type predicted for the Bloom (1956) Taxonomy, the matrix of intercorrelations should exhibit distinctive features. One feature is that the correlations along any row of the matrix should have successive magnitudes similar to a permutation of those along any other row. Secondly, small-magnitude partial correlations between non-successive levels tests should be observed when the effects of tests at intermediate levels are removed. And furthermore, the plotting of the two characteristics of smallest-space analysis should produce concentric figures with the most fundamental levels most central. Stoker and Kropp (1971) found limited support for a hierarchy involving the taxonomic levels with the exception of the knowledge level. Tests at the knowledge level did not form part of the concentric pattern suggesting lack of hierarchical relationship with the other tests.

The studies of the Bloom (1956) Taxonomy are not convincing on the question of whether or not application stems are really part of a hierarchy. While the study of Smith (1970) showed application to be in a hierarchy after knowledge and comprehension, Stoker and Kropp (1971) found that knowledge was not part of the hierarchy. Poole (1971) was not able to substantiate the hierarchy and noted severe difficulties in obtaining consistent judgements on the classification of items. Except for Poole (1971) all of the cited studies did find that comprehension and analysis formed at least a limited hierarchy.

Avital and Shettleworth (1968) reasoned that comprehension and application differ not in complexity of process but in the degree of novelty of the situation in which knowledge is combined. They argued that the knowledge level involved recall or recognition, and that the analysis-synthesis-evaluation level required the student to produce a solution entirely new to him, but the comprehension and application levels required some elements of recall and some elements of novelty. Since the exact proportions are difficult to specify, comprehension and application were combined as algorithmic thinking. Avital and Shettleworth's classification scheme was derived for mathematics items, but the mathematical nature of many physics items suggests that the scheme may be applicable.

The defining of low context test items in terms of the presence of non-relevant data and multi-concept response alternatives provides an operational definition of low versus high context items. Thus the classifying of items as low or high in context is much less a matter of judgement than is the classification of items within the Bloom (1956) categories.

Summary

Low context test items have been shown to be more closely related to application as defined by Hurd (1973) than to application as assessed in the study of Smith, Tyler et al. (1942). Depending upon the experience of the student low context items could be either knowledge or application in their level of assessment. In spite of the presence of non-relevant data and multi-concept response alternatives those students who use recall or recognition to determine the response are operating at the knowledge level.

The studies of Poole (1971) and Smith (1970) have suggested that attempts to validate a hierarchy of the Bloom (1956) Taxonomy may be foundering on the lack of consistent judgements on the placement in the hierarchy of items, particularly items beyond the knowledge level.

The rating of items as low or high in context is based on operational methods likely to have less ambiguity than those used to rate items in the Taxonomy (Bloom, 1956).

CHAPTER 3

INSTRUMENTATION AND DESIGN

Assessment context is the criterion variable of the present study. A full explanation of assessment context and the tests constructed to measure it form an important part of the first section of the present chapter. In addition, the tests used to appraise the independent variables are described. The second part of the chapter embodies the details of the design of the study including sampling, test administration and data analysis procedures.

In the present investigation relationships were sought among assessment context, physics achievement, extent of field independence, breadth of categorization, and numerical and verbal abilities. Assessment context was the criterion variable and was assessed by three tests: the High Context Test (HCT), the Low context Test (LCT) and the Discrepancy Detection Test (DDT).

The HCT and the LCT were both verbal tests with multiple responses. For reasons described below the HCT and LCT were combined for administration and designated as the Physics Test (PT). The DDT, also low in assessment context, consisted of diagrammatic items with free-response format.

Physics achievement was measured by a multiple choice test having two sub-scales indicating achievement at the knowledge and the algorithmic thinking levels. The Physics Achievement Test (P30) must not be confused with the PT which is the combined form of the HCT and LCT. The two sub-

6

scales of the P30 were named the Avital-Shettleworth Knowledge Test (ASK) and the Avital-Shettleworth Algorithmic Thinking Test (ASAT).

Extent of field independence and breadth of categorization were assessed by means of the Hidden Figures Test (HFT) (Jackson et al., 1964) and the Category Width Scale (CWS) (Pettigrew, 1958).

General scholastic ability in the verbal and quantitative realms were measured by the Cooperative School and College Ability Tests, Verbal (SCAT-V) and Quantitative (SCAT-Q).

In the first subsection which follows the meaning of assessment context and the methods of rating the assessment context are described. Succeeding subsections describe the construction of the assessment context tests and the properties of the tests used to measure the independent variables of the study.

Assessment Context

The assessment context value of a multiple choice test item is defined in terms of the following properties: (a) the amount of redundancy in information used for correctly solving the item, (b) the presence of information in the item stem which is irrelevant to solving the item, and (c) the diversity of the concepts to be found among the various alternative responses provided in the item.

Redundancy of the essential information in a test item is evident when a diagram is used to repeat information given verbally, or when excess information is provided by repetition of concepts by symbols or by naming the originator of a law of physics in addition to stating the law, or by providing measurement unit names as well as the names of the quantities being measured. Consider the following contrasting examples:

A. Newton's second law states that a mass of (m)kg will accelerate at (a) m/sec^2 when acted upon by an accelerating force of (f) newtons. If the force is increased to (2f) newtons, and the mass is decreased to (1/2 m)kg, the acceleration will become

1. (4a) m/sec^2
2. (2a) m/sec^2
3. (a) m/sec^2
4. (1/2a) m/sec^2
5. (1/4a) m/sec^2

B. When a mass is acted upon by an unbalanced force, it will accelerate. The magnitude of the acceleration is directly proportional to

1. the mass and to the force
2. the force and to the square of the mass
3. the mass and inversely proportional to the force
4. the product of mass and force
5. the force and inversely to the mass

Essentially the same information is conveyed in A and in B but with less redundancy in B. In Item A the principle required for solving the item, Newton's second law, is named; also the usual symbols for mass, acceleration, and force are presented as well as the physical units in which each is measured. On the other hand, item B contains only a non-quantitative statement in the item stem about mass, force, and acceleration. Item A is rated higher in context value than item B.

The second element which affects context value is the presence of information irrelevant to that required for solving the item. When irrelevant information is present in an item, the item is rated as being low in context value. The following items illustrate the presence and absence of irrelevant information, respectively:

C. A wooden cube has a mass of 1.0 slug. The cube is 1.2 ft on each edge; its specific gravity is 0.29. An unbalanced force of 10 lb accelerates the cube through a distance of 20 ft, starting from rest. The acceleration of the cube is

   1. 24 ft/sec^2
   2. 10 ft/sec^2
   3. 0.31 ft/sec^2
   4. 200 ft/sec^2

D. A mass of 1.0 slug is acted upon by an unbalanced force of 10 lb. The acceleration of the object is

1. 100 ft/sec^2
2. 10 ft/sec^2
3. 0.10 ft/sec^2
4. None of the above answers

Item C contains information which is irrelevant to solving the item. Data on the size of the object, the specific gravity, the distance through which the force acts, and the starting speed are unnecessary for calculating the acceleration. On the other hand, item D contains only information which is needed for answering the item. The information in test item D is redundant in that the specific units do not actually need to be provided for mass, force or acceleration. The mass could have been designated in "mass units", or "British units of mass" rather than in slugs. The provision of "slugs", "newtons" and "m/sec^2" provides the student with extra assistance for recalling how to solve this item. The student may have forgotten how to calculate acceleration from mass and net force, but naming of the mass unit, the slug, could help him recall $f = ma$, since this is usually the relationship in which physics students are first called upon to work in specific mass units. Recalling $f = ma$ probably makes the item solvable. Although the provision of the particular measurement units represents redundant information it is not irrelevant information as is included in item C.

The third property on which item context depends is the diversity of physics concepts presented among the response alternatives to the item. If all of the responses to the item are related to the same physics concept, for example, all kinetic energy amounts, the item is of higher context value than if each of the alternative responses demands the consideration of different concepts. By the definition of physics concept adopted here kinetic energy, ft/sec^2, acceleration are three different concepts whereas 2 ft/sec^2 and 3 ft/sec^2 are assumed to be the same physics concept. The following two items have, respectively, single concept and multiple concept response alternatives.

E. Two objects, one having a mass (m), the other a mass (2m), are simultaneously released from the top edge of a tall building. By the time the two masses have fallen 50 ft, the larger mass will have

1. a kinetic energy 4 times that of the smaller
2. a kinetic energy the same as that of the smaller
3. a kinetic energy twice that of the smaller
4. a kinetic energy one-fourth of that of the smaller

F. Two objects, one having a mass (m) and density (d_1), the other a mass (2m) and density (d_2), are simultaneously released from the top edge of a 100 ft high building. By the time the two masses have fallen 50 ft, the larger mass will have:

1. the same kinetic energy as the smaller mass
2. the same momentum as the smaller mass
3. the same net force on it as the smaller mass
4. the same velocity as the smaller mass.

Because of the wider diversity of concepts in the alternative responses to item F as compared with item E, item F would be rated lower in assessment context than item E.

The assessment context of an item, then, is related to:

(a) the amount of redundancy of essential information,

(b) the amount of information in the item stem which is irrelevant to answering the item,

(c) the extent of the diversity of physics concepts presented among the response alternatives.

An item having redundancy of the essential information, no irrelevant information, and only one physics concept among the response alternatives is rated high in assessment context. But what of items having only one or two of these properties?

It is possible to write test items with all possible combinations of the three properties described above. The number of combinations of these three properties in a test item is eight, based upon redundancy of the essential information (Yes or No), only relevant information (Yes or No), and single concept response alternatives (Yes or No). Thus the number of possible combinations is 2 x 2 x 2 = 8 and the assessment context of an item need not be nominal in its level of measurement. The assessment context of an item varies along an ordinal scale depending upon the particular combination of the three properties above which the item possesses.

In order to accentuate the magnitude of relationships existing among assessment context and the other variables of the present study, items were written which could be rated as high or low in assessment context because they did not contain the "middle" combinations of the three assessment context criteria. This meant that items high in assessment context had to have redundancy of the essential information, no irrelevant data, and a single physics concept for the response alternatives. In contrast, items low in assessment context had to have two or more physics concepts among the response alternatives and at least one example of irrelevant data. However, the items might or might not have contained redundancy of essential information. Attempts to write items without redundancy of the essential information were not successful since the English language itself is redundant in the sense that certain words may be deleted from sentences without the sense of the sentence being lost (Attneave, 1958). The redundancies of essential information in the low context items were minimized by not providing diagrams in addition to verbal statements and by not using standard physical units for any of the data provided.

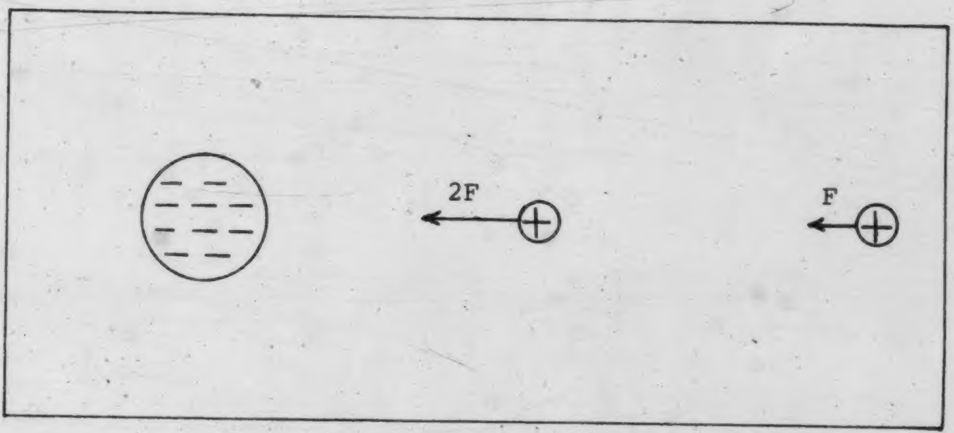
Besides the verbal items described above, a set of diagrammatic items was also prepared. Each item contained a diagram and sometimes a few labels of parts or words of description or assumption. Students responding to the items were told that the diagram might or might not portray a situation in which a law of physics was violated. If the student felt that there was a contradiction between what he knew of physics and what was shown in the diagram he was to explain the nature of the contradiction; if not, he was to say that the diagram was all right. Because of the nature of the task, the diagrammatic items eventually formed a Discrepancy Detection Test (DDT). An example of a DDT item is presented in Figure 2.

The DDT items are low in assessment context. Even the simplest diagrammatic portrayal of a situation contains data which are irrelevant to securing the required solution. In Figure 2 the negatively charged object is shown to be larger than either of the positively charged objects and shown to have 10 times as many negative charges as there are positives on either of the other objects. Yet these features are irrelevant since the basic nature of the item would not be changed if the negatively charged object were twice as large or half as large or if it had 10 times as much negative charge or half as much as is shown. Thus the size of the objects and the exact amount of the negative charge are not relevant data.

The DDT items contain little if any redundancy of the

Figure 2

A Diagrammatic Item from the Field of Electrostatics



essential information. In Figure 2 the essential information is comprised of the equivalent amounts of positive charge on the two right hand objects, the equal spacings along a line joining the three objects, the force amounts (2F) and (F), and the direction of the forces on the right hand objects. Altering any of these data would alter the situation possibly changing the nature of the item from one in which there is a discrepancy to one in which there is not.

The DDT items potentially suggest several physics concepts for their solution. In solving Figure 2 the student may consider a gravitational force explanation, an explanation in terms of the direction of forces among like and unlike electrical charges, as well as Coulomb's Law for electrostatic forces.

The DDT items are then considered low in assessment context since they contain irrelevant data, have minimal redundancy of the essential information, and, potentially at least, give rise to alternative physics concepts for their solution.

There is a discrepancy in Figure 2. Since electrical forces are governed by an inverse square law the right handed positively charged object should experience one-quarter as much electrical force as the similarly charged objects located half as far from the attracting charge; instead the right handed object is shown to experience one-half as much force as the other positive charge.

The construction of the criterion tests is described next. Because of the HCT and the LCT were both verbal, multiple choice tests, and because they were eventually combined for administration as a single test, their items were not distinguished at many of the test construction stages. The DDT, because it differed in item form and response format, was not constructed in exactly the same way as the LCT and HCT.

The LCT and HCT. Because the investigator's teaching competence was strong in physics, the items for the HCT and LCT were written for the content of the Alberta Physics 30 course based upon the textbook _Fundamentals of Physics_ (Stollberg and Hill, 1968) as used in the 1971-72 school year. Physics 30 consists of approximately equal-time allotments for statics, kinematics, and dynamics under mechanics and for electrostatics and magnetostatics, current electricity, and electromagnetism under electricity and magnetism.

Forty-five items low in assessment context and 45 items high in assessment context were written by the investigator with approximately equal distribution of items among the various content areas. The procedure adopted was to write alternately an item high in assessment context and then one low in assessment context until each of the six main areas of the Physics 30 course was represented with 15 items. Although items both high and low in assessment context were not written for each concept or principle of physics assessed, approximately equal numbers of high and low assessment context were constructed for each content area of the course.

Because of the novel features of the low context items, a preliminary administration of 15 randomly selected items was carried out with 11 senior high school students who had recently completed the Physics 30 course from the Stollberg and Hill (1968) textbook. The purpose of the tryout was to assess the adequacy of the directions for the LCT, the time requirement for the items, and the reactions of students to the items. The results indicated that the directions were suitable, that a time of about 1 1/4 minutes per item was ample, and that students did not react unfavorably to the low context items.

Subsequently a pretest of all 90 items was carried out with four class sections of introductory physics, Physics 100, at the University of Alberta. Physics 100 whose contents are mechanics, electricity and magnetism is equivalent to Physics 30 for the Province of Alberta. The students of Physics 100 are not required to have completed any other course in physics. Persons familiar both with Alberta High Schools and the University of Alberta are of the opinion that Physics 30 students, who have elected to enrol in a course with a reputation for difficulty, are generally superior in science ability to the students of

Physics 100, who are often required to enrol in the course to achieve a prerequisite for some course in which they are interested. Nevertheless, the assumption was made that the capabilities of Physics 30 and Physics 100 students were comparable.

Various time allocations were specified by the instructors of the four university class sections. Items were selected randomly without replacement from the pool of 90 items for four subsets of 21, 21, 18, and 30 items respectively. Thirty minutes of class time were available for each of the 21-item subtests, 50 minutes each for the 18-item and 30-item subtests. The 50 minute period was needed for the 18-item subtest because an additional set of diagrammatic items was also administered. The numbers of students writing each subtest were as follows:

first 21-item subtest,	27 students
second 21-item subtest,	13 students
18-item subtest,	33 students
30-item subtest,	45 students.

For each subset the item difficulties were calculated. Items for which the difficulties were less than 0.25 or greater than 0.80 were eliminated from further consideration, with four exceptions. The four exceptions were items for which the teacher of the university class felt that the topics were inadequately treated and for which the investigator considered there was ample treatment in the Physics 30 course. After elimination of items of inadequate difficulties, biserial correlations between item and test total were calculated along with the cluster analysis procedure of Loevinger, Gleser and DuBois (1953) for each subtest.

A panel of three experienced physics teachers rated the items not eliminated by the difficulty criterion as to assessment context level, and as to content validity. The instructions to the judges are presented in Appendix A.

The criteria for retaining items were: difficulty levels between 0.25 and 0.75, either high or low on assessment context, majority agreement of judges on content validity, biserial correlations with subtest totals greater than 0.30 and, membership in a cluster based on the cluster analysis procedure (Loevinger et al., 1953). Given the small number of students who wrote the subtests the biserial correlation and cluster membership criteria were relaxed for some items in order to have 20 items low in assessment context and 20 items high in context with sampling from each content area for both tests. Item data based upon pretests and pre-judgements are summarized in Table 2.

Items 4, 19, 34 and 38 (Appendix B) did not meet the difficulty criterion as described above. Items 11, 25, 29 and 31 (Appendix B) were revised in order to meet content validity and assessment context criteria. Hence the pretest data no longer applied to these items.

The 40 assembled items constituted a test called the

Table 2

Summary of the Pretest Data for the PT Which Includes the HCT and the LCT as Subtests

Item[a]	Assessment context	Difficulty	Biserial r	Dominant cluster membership
1	High	0.75	-0.15	Yes
2	Low	0.38	-0.05	No
3	Low	0.24	0.31	No
4	High	0.15	0.55	No
5	High	0.31	0.25	Yes
6	Low	0.29	0.12	No
7	High	0.23	0.57	Yes
8	Low	0.85	0	Yes
9	Low	0.51	0.44	No
10	Low	0.42	0.52	Yes
11[b]	High	-	-	-
12	Low	0.63	0.03	Yes
13	High	0.42	0.09	No
14	High	0.42	0.24	Yes
15	Low	0.39	0.39	Yes
16	High	0.55	0.28	Yes
17	Low	0.44	0.40	Yes
18	High	0.67	0.60	No
19	Low	0.16	0	No
20	High	0.24	0.31	Yes

[a] All items were certified as to content validity.

[b] This item later was revised. Its content validity and assessment context were re-rated, but not difficulty, biserial correlation coefficient or cluster membership.

Table 2 (continued)

Summary of the Pretest Data for the PT Which Includes the HCT and the LCT as Subtests

Item	Assessment context	Difficulty	Biserial r	Dominant cluster membership
21	High	0.15	0.53	No
22	Low	0.45	0	Yes
23	High	0.33	0.26	Yes
24	High	0.27	0.39	Yes
25[b]	Low	-	-	-
26	High	0.07	0.69	No
27	Low	0.27	0.30	Yes
28	Low	0.63	-0.21	Yes
29[b]	Low	-	-	-
30	High	0.48	0.40	No
31[b]	High	-	-	-
32	Low	0.24	0	Yes
33	Low	0.15	-0.12	No
34	High	0.11	0.44	No
35	High	0.62	0.32	No
36	High	0.38	0.30	Yes
37	High	0.38	-0.61	Yes
38	Low	0.13	0.13	No
39	High	0.40	0.39	Yes
40	Low	0.46	0.26	No

[b] This item was revised. Its content validity and assessment context were re-rated, but not difficulty, biserial correlation coefficient or cluster membership.

Physics Test which is presented in Appendix B. The subset of 20 items low in assessment context formed the LCT and the 20 items high in assessment context the HCT.

Research related to the HCT and LCT. The test items of the LCT exhibit several properties which may be called non-standard. The LCT items contain multiple physics concepts among the response alternatives. The effect of the multiple concepts is to make some responses longer than others in the same item, and for some items the verb of the sentence connecting the item stem and response alternatives was located in the response alternative rather than in the item stem. The non-standard practices just noted conflict with Nedelsky's (1965, p. 166) recommendation that response alternatives should be homogeneous in form. Yet studies of Dressel and Schmid (1953), Dunn and Goldstein (1959) and McMorris, Brown, Snyder and Prezek (1972) suggest that the impact of such non-standard practices is minimal.

Dressel and Schmid (1953) designed response methods which forced college physical science students to evaluate each alternative rather than to pick the one suggested first by analysis of the item. They reported no significant differences among student performances on various standard and non-standard response formats and concluded that considerable liberty could be taken with modifications to the multiple choice item without impairing its efficiency (Dressel and Schmid, 1953, p. 595).

Dunn and Goldstein (1959) and McMorris et al. (1972) investigated the effect of item faults including the repeating of words in stem and response alternatives, unequal response lengths, and grammatical inconsistencies between stem and response alternatives. The results for the army subjects of Dunn and Goldstein and the senior high school subjects of McMorris et al. were similar in that the stem faults tended to make items somewhat easier, but the validities and reliabilities were not significantly affected. The findings reenforce the conclusion of Dressel and Schmid (1953) that multiple choice items can be modified without serious consequences.

The presence of irrelevant information in items of the LCT can be construed as similar to "window dressing" which Ebel (1972, p. 217) warns against. Board and Whitney (1972) introduced window dressing into multiple choice items with political science content. The subjects of the study were college undergraduates. While the difficulty levels were not affected by the window dressing, Board and Whitney (1972) found that the validity of a test having such items was somewhat lower than a standard test and the reliability was significantly less.

The evidence suggests that the impact of the non-standard properties of items on the technical quality of the LCT is likely to be minimal, except possibly that the reliability may be somewhat less than that of a similar test having standard items. The possibility of lower reliability than that possible with standard items is accepted for the present study since the LCT is a research instrument and not one being used for decision making on student achievement.

The DDT. Forty-five items were prepared initially. Approximately equal numbers of items were constructed for each of the major areas of the Alberta Physics 30 Course, 1971-72, based on the textbook by Stollberg and Hill (1968). The physics areas were statics, kinematics, dynamics, electrostatics, magnetostatics, current electricity and electromagnetism. The items were randomly assembled to form a test and the test was administered to 11 high school seniors who had recently completed the Physics 30 course.

The result of the preliminary administration showed that the instructions to students required improvement in order that scoring be reliable. The responses of the 11 students were sometimes unclear as to the contradictions they felt existed between their knowledge of physics and that which was portrayed in the diagrams. Some students named only the principle violated; others stated only what they felt was incorrect in the item without naming the principle violated. Subsequently the directions were modified asking subjects to look for contradictions between what was pictured and that which their knowledge of physics led them to expect. If a contradiction was detected, both the rule of physics violated and the specific violation were to be stated. If no discrepancy existed the student simply put the letter "C", for correct, on his response paper beside the number of the item. The time required for students to respond to the DDT items averaged about 3/4 minutes per item.

The items with revised instructions were then administered to one class section of Physics 100 at the University of Alberta. Thirty-three students responded to the items. The difficulty levels and biserial correlations between item performances and test totals were calculated.

The decision had been made by the investigator to restrict each of the criterion tests to 20 items because of restricted time availability of the students for the study. Items for the DDT were selected by the following criteria: 1) content coverage of each of the eight physics areas, 2) item difficulties in the range 0.20 to 0.80, 3) biserial correlations of 0.30 or greater. Pretest data for the 20 items retained for the DDT are summarized in Table 3. The four items not meeting the difficulty criterion were retained because there was reason for the investigator to believe that they related to concepts not stressed in the physics course of the pretest group but emphasized in the course of study of the target population.

The DDT is presented in Appendix C.

Review of studies related to the DDT. A method of assessing physics achievement using pictorial items has been developed by Podrasky (1971). Podrasky felt that the verbal

Table 3

Summary of the Pretest

Data for the DDT

Item	Difficulty[a]	Biserial r	Dominant cluster membership
1	0.27	0.28	No
2	0.58	0.10	Yes
3	0.70	-0.09	No
4	0.20	0.31	No
5	0.10	0.44	Yes
6	0.21	0.57	Yes
7	0.79	0.28	Yes
8[b]	*	*	*
9	0.20	0.44	Yes
10	0.36	0.27	Yes
11	0.64	0	No
12	0.68	0.70	Yes
13	0.27	0.38	Yes
14	0.10	0.19	No
15	0.21	0.90	Yes
16	0.15	0.87	Yes
17	0.61	0.28	Yes
18	0.21	-0.20	No
19	0.03	–	No
20	0.33	0.67	Yes

[a] N=33

[b] Subsequently this item was revised in a major way; therefore the pre-test data do not apply.

nature of most physics tests might penalize students of poor verbal ability and endeavored to prepare a multiple choice test with no verbal component. Each item of Podrasky's (1971) test consisted of two transparent slides which were projected on two separate screens. On one screen was a picture or diagram depicting a situation in which a principle or phenomenon of physics is evident. A label naming the principle or phenomenon appears on the picture. On the second screen are projected four different images one of which contains a situation in which the principle or phenomenon is relevant. The student chooses the one picture on the second screen which seems most closely related to the stimulus picture. For example, one of the stimulus slides shows a cross-section of an air foil with flow lines around it and the words "Bernoulli's Principle" below. Shown in the four response pictures are a dirigible, an air-cushion vehicle, a hydrofoil boat and kite. No words are provided with the response alternatives. The hydrofoil boat is the best response for the item. Podrasky (1971) administered a conventional physics test and a verbal ability measure as well as the 73-item pictorial test to his sample of 226 high school physics students. An equation was calculated for predicting achievement on the pictorial test from scores on the regular test. Podrasky then compared actual and predicted scores and found that about 55 percent of those students achieving better scores than predicted on the pictorial test scored below the mean on the verbal ability measure. Podrasky (1971) interpreted the results to mean that about 10 percent of students benefitted by a non-verbal rather than a verbal assessment of physics achievement. Since Podrasky (1971) did not indicate the extent to which the 10 percent of students benefitted, his results are difficult to assess. They suggest however that in the present study a weaker relationship may exist between the DDT and the SCAT-V than that between the LCT or HCT and the SCAT-V.

Another approach to ascertaining the abilities associated with responding to various types of physics test items has been followed by George (1971) who prepared a physics test which included subtests varying in verbal and diagrammatic content. One subtest was completely verbal. At the other extreme was a subtest with diagrammatic items although the diagrammatic subtest was not completely non-verbal since many of the items contained a sentence or two of direction or explanation. The subjects of the study, senior high school students enrolled in a first course in physics, were administered the Differential Aptitude Tests in addition to the physics tests.

George (1971) reported that while the intercorrelations among the physics subtests were all positive those involving the diagrammatic subtests were less positive than the others. While verbal ability was the best predictor of success in other subtests, mechanical reasoning was the only significant predictor of success on the diagrammatic test. George's (1971) study implies that the ability to do diagrammatic items is somewhat different from that required for verbal items. The diagrammatic items of George (1971), like those of the DDT, sometimes have a few words of explanation; however, the diagrammatic items are multiple choice not free response as in the DDT. In spite of the differing response format the DDT can be expected to be less strongly related to the SCAT-V than either the LCT or HCT.

Cognitive Style Variables

The cognitive style variables of the present study are extent of field independence and breadth of categorization. Chosen to assess the variables were the HFT (French et al., 1963) and the CWS (Pettigrew, 1958). Each of the tests is described and related literature reviewed in the subsections which follow.

The HFT. The HFT is a group test of the pencil-and-paper variety designed to assess extent of field independence (Jackson et al., 1964). The test is moderately speeded with 10 minutes of time permitted for each 16-item half of the test. At the top of each page of the test booklet is the same set of five simple geometric figures labelled A-E. Below each set are a series of complex figures with the letters A-E below each one. Each complex figure represents an item. Within each complex figure is one of the five simple figures of the same size and orientation as it is given at the top of the page. In order to respond to the item, the test subject must find which one of the five simple figures is located within the complex figure found. One point is allotted for each item in which the simple figure is correctly identified. Thus the maximum possible score is 32.

Internal consistency reliabilities for the HFT, possibly spuriously high due to the speeded nature of the test, have been reported as 0.71 (Jackson et al., 1964) and 0.79 (Boersma, 1968). Jackson et al. (1964) have reported a split-half reliability of 0.63 for a 10-week interval between test administrations while Fleishman and Dusek (1971) have presented a test-retest reliability of 0.72 with test administrations separated by a few hours.

The CWS. The test used in the present study to assess breadth of categorization is the CWS. Following the practice of Steiner and Johnson (1964), Mascaro (1968) and Eagly (1969), the 10 items with the highest validity of the original 20 (Pettigrew, 1958) constituted the CWS. The test is not speeded.

Each item of the CWS presents an average value of some varying quantity. In part (a) of the item the subject is asked to select which one of four values, each one larger than the given average, is probably the largest quantity of the property likely to be found. In part (b) the subject is required to select the smallest quantity of the property likely to be found from among four values all smaller than the given average value. To illustrate, the first item of the CWS is presented below:

> It has been estimated that the average width of windows is 34 inches. What do you think:
>
> (a) is the width of the widest window . . .
>
> 1. 1363 inches ( ) 3. 48 inches ( )
>
> 2. 341 inches ( ) 4. 81 inches ( )
>
> (b) is the width of the narrowest window . . .
>
> 1. 3 inches ( ) 3. 11 inches ( )
>
> 3. 18 inches ( ) 4. 1 inch ( )

In scoring the CWS the sum of the weighted choices is calculated. The weightings which are not made known to the test subject provide for 0, 1, 2, or 3 for each part of each item depending on the extent of the distance of each alternative choice from the average value which is stated in each item. Values farthest from the mean are weighted most heavily. The CWS contains 10 items, and thus there are 20 parts to the test. The maximum possible score is 60 when the maximally weighted choice is selected for each of the 20 parts.

Odd-even reliabilities of 0.86 and 0.93 have been reported for the original 20-item scale (Pettigrew, 1958). Eagly (1969) calculated a corrected split-half reliability of 0.80 for the CWS.

Ability Variables

Verbal and numerical abilities were assessed for the students participating in the present study. The variables were measured by the SCAT-V and SCAT-Q tests. The properties of the tests are described in the following subsection.

SCAT-V and -Q. As part of a province-wide testing project in Alberta the SCAT was administered to most subjects of the present study during the spring term of their grade nine year, three years prior to the present study. Form 3A of the SCAT was administered at that time. The SCAT-V and SCAT-Q are moderately speeded tests of 60 and 50 items respectively. Each test requires 50 minutes for administration. The response formats are multiple choice.

Green, writing in Buros (1965), states that the SCAT are academic aptitude tests which correlate well with general intelligence, and quotes correlations of 0.84 between WAIS and SCAT, 0.88 between WAIS-V and SCAT, 0.77 and 0.81 between the OTIS Quick Scoring Mental Ability Test and SCAT for two large samples of junior college subjects in support of the statement. For the present study the SCAT is assumed to be an adequate measure of verbal and numerical ability so that these variables may be controlled in examining relationships of the other independent variables and the criterion variables.

The sample for the present study is composed of 12th grade students who wrote the SCAT, Level 3, three years earlier. Tully and Hall (1965) have reported test-retest reliabilities for 100 high school students over a one-year period from 9th to 10th grade; the reliabilities ranged from 0.86 for the SCAT-Q to 0.93 for total scores. The test-retest reliabilities compare favorably with the internal consistency reliabilities in the neighborhood of 0.94 (SCAT Technical Manual, 1957). The internal consistency reliabilities may be somewhat high because of the moderately speeded nature of the tests (Green, 1965). The tests appear to be quite stable over time, at least for a one-year interval.

Indications of the predictability of high school subject grades are obtained from studies of Tempero and Ivanoff (1960) and from Sommerfield and Tracy (1963). Tempero and Ivanoff (1960) reported correlations of 0.67 and 0.56 for SCAT-Q with algebra achievement and physics achievement, respectively. The numbers of high school students for each subject field were 70 and 61, and the time interval was one year between administration of SCAT and the determination of achievement in the subject field. For algebra and physics with SCAT-V the correlations were 0.26 and 0.51, respectively. Sommerfield and Tracy administered the SCAT to 152 ninth grade students and correlated the test performance with algebra achievement one year later. The correlation of algebra with each of SCAT-Q, SCAT-V and SCAT-Total were 0.52, 0.42, and 0.53. The results indicate that for algebra and physics the SCAT-Q is a better predictor than the SCAT-V. For predicting achievement on the HCT, LCT and DDT over a three year period the SCAT tests can be expected to be less efficient than over the one year period in the above studies, and probably cannot predict more than 16 percent of the variance of the HCT, LCT and DDT.

Physics Achievement Variables

Although only one test was administered, three subscores were derived from it to assess performance on the three Avital and Shettleworth (1968) categories.

Avital and Shettleworth define the knowledge category as equivalent to the knowledge level of the Bloom (1956) Taxonomy. Hence, recall or recognition are the identifying features. The algorithmic thinking level (Avital and Shettleworth, 1968) encompasses the comprehension and application levels of the Taxonomy. Algorithmic thinking is characterized by the use of symbolic process or multistep procedures involving elements of novelty. The third category of the Avital and Shettleworth (1968) scheme of objectives is called open search. Open search demands insight into a new conceptualization of understanding. As such, open search is difficult to assess within the limitations of a confined test. Although Avital and Shettleworth (1968, chap. 4) have pictured open search as encompassing the Taxonomy categories of synthesis and analysis, the strong emphasis on insight for open search rather than on the Hierarchy of complexity of the Bloom (1956) Taxonomy suggests that open search and analysis-synthesis are not parallel.

The requirements of open search test items are such that the likelihood of finding them on a multiple choice test administered within a specified time limit were remote. Therefore, the variable of open search physics achievement was not assessed. In the subsection which follows the construction of the Physics Achievement Test (P30) is discussed as well as the procedures followed in determining which items of the P30 assessed physics achievement at the knowledge and the algorithmic thinking levels.

The ASK and ASAT subtests. All students of Physics 30 in Alberta, including the students in the present study, were administered the P30 at the conclusion of the course. The P30 met the construction requirements of all senior achievement tests prepared by the Alberta Department of Education (Wood et al., 1968). The requirements are as follows:

1) Approximately 60 items (The P30 had 65)
2) Items written by subject teachers
3) Objective items with 4-choice responses, one to be selected
4) Items representing various levels of Bloom (1956) Taxonomy
5) Item pretests with at least 100 students of the target population
6) Biserial correlations of item and test total of 0.30 or greater

7) Item reliabilities of 0.15 or greater

8) Item revisions by subject teachers.

The students of the present study wrote the P30 of June 1972. The test has not been reproduced in the present study at the request of the Alberta Department of Education.

Although the items of the P30 were rated as to Taxonomy level by the item writers, the decision was made to secure another judgement on item categorization according to the Avital and Shettleworth scheme. Accordingly, a panel of four judges rated the items of the P30 as: 1) knowledge level, 2) comprehension or application level, or 3) higher levels of the Bloom Taxonomy. The judges were two university professors of science education and two graduate students in science education, one of whom was the investigator. Both of the graduate students had taught senior high physics courses and one had taught the Physics 30 course.

The judging proceeded as follows. Independently each judge classified each of the 65 items. Then the panel of judges met together to discuss those items for which there was not unanimous agreement as to categorization. Twenty such items were discussed. If unanimity of agreement was not achieved by the discussion then the majority decision prevailed. If no majority emerged for an item, the item was not considered in the analysis. One item was rejected.

The item judging yielded the following results:

knowledge level	8 items

algorithmic thinking level	48 items
higher taxonomic levels	8 items
not classifiable	1 item.

The investigator also applied the criteria for determining item membership in the HCT and LCT. Forty-four of the P30 items were candidates for membership in the HCT, 3 for the LCT, and 18 were intermediate between HCT and LCT. By level the assessment context ratings were distributed as follows:

knowledge level:	2 HCT,	4 LCT,	2 neither
algorithmic thinking:	37 HCT,	11 LCT,	0 neither
higher levels:	4 HCT,	3 LCT,	1 neither
unclassified:	1 HCT		

The ASK was an eight-item subtest of the P30, and the ASAT a 48-item subtest of the P30. For reasons already explained no assessment was made of physics achievement at the open search level.

Design

The details of the sample, of the administration of tests, and of the mathematical procedures applied to the test scores are presented in the subsections which follow.

The Sample

The population of students consisted of senior high school physics students. More specifically, these were the students who wrote the June 1972 Alberta Department of Education examination for the Physics 30 course based upon the textbook by Stollberg and Hill (1968). The sample was made up of students writing the P30 at three Edmonton high schools, two schools from the Edmonton Public System and one from the Edmonton Separate System.

In order to secure the sample 10 Edmonton high schools were contacted for possible participation in the study. Two of the schools had students enrolled in an alternative physics course and were thus disqualified. Five schools did not wish to participate. The participating schools included an Edmonton Public school located near the centre of the city, an Edmonton Public School in a new area near the edge of the city, and an Edmonton Separate school which draws students from both new and old regions of the city. The make-up of the sample is presented in Table 4.

Physics teachers at the downtown school, designated Public 1, indicated that five students in the sample were having difficulty in reading English because of foreign-language backgrounds; furthermore, eight students including two of those with English-language difficulties were enrolled in a special retraining program for those returning to high school after having withdrawn more than a year previously. Teachers at the other two sample schools noted no such special circumstances for any of their students.

Test Administration

Since the contents of the HCT, the LCT, the DDT and the P30 were based on physics, the testing had to be delayed

TABLE 4

The Sample by School and by Sex

School	Sex	
	M	F
Public 1	55	12
Public 2	39	8
Separate	28	2
Totals	122	22

until near the end of the school term when the Physics 30 course was nearly completed and yet early enough to avoid the declining attendance of students which occurs near the end of classes. The LCT and HCT in combined form were administered first in a 55-minute period followed several days later by the DDT, the HFT and the CWS in a second 55-minute period during the third and fourth weeks of May. The P30 was administered about three weeks later during the schools' regular examination period.

The HCT, LCT, DDT, HFT and CWS were administered by the investigator except in the Public 1 school where, because of two simultaneous Physics 30 classes, the Physics Department Head administered the tests to one class of students. The P30 was administered by the classroom teachers.

In order to justify the use of the 110 minutes of student testing time the students' answer sheets were scored and checked by the investigator within two school days of administration in each school. The answer sheets were then returned to the students in order to maximize the value of the HCT, LCT and DDT as devices for reviewing the content of the physics course. Each teacher retained a copy of the HCT, LCT, DDT and an answer key to each.

Mathematical Procedures

Students responded to the HCT and LCT by blanking out appropriate space on an IBM General Purpose Score Sheet. Responses to the HFT and CWS were made on the answer sheet.

Each of the tests was hand-scored and totalled. Then each student's responses were entered into a computer via a typewriter terminal and again scored electronically by comparison with an answer vector. The procedure adopted provided a check on the accuracy of the hand scoring and the computer scoring. Subsequently the computer-stored data were transferred to cards for further analysis.

The free-responses to the DDT were read and scored as one or zero for correct or incorrect. After rechecking, the item results were transferred to computer and again totalled to check on recording errors.

Raw scores for the SCAT-V and SCAT-Q were obtained from the records of the Provincial Department of Education. Xerox copies of the response sheets for the P30 were obtained from the Department of Education for the three schools which participated in the study.

The DDT, HCT, LCT, HFT, CWS were analyzed for item difficulty and internal consistency reliability. Intercorrelations were calculated among all measures.

Detailed descriptions of the multiple linear regression and canonical correlation calculations which were made are described in the chapter dealing with hypotheses and statistical procedures.

CHAPTER 4

HYPOTHESES AND DATA ANALYSIS PROCEDURES

In the present chapter the hypotheses of the study are developed in detail, following which the data analysis procedures for testing the hypotheses are described.

Hypotheses

The criterion variables of the present study consist of physics ability as assessed by a verbal, multiple choice test high in assessment context, a verbal, multiple choice test low in assessment context, and a diagrammatic test consisting of discrepancy detection items. The predictor variables consist of verbal and quantitative ability, physics achievement, extent of field independence, and breadth of categorization. A number of multivariable hypotheses are presented in order to assess relationships which may exist. The order of presentation of the hypotheses is the same as the order in which they are to be tested.

In multivariate studies, Walberg (1971) advocates the testing first of overall relationships among the variables by a multivariate test; hence the first hypothesis presented will involve relationships among both sets of variables. After the overall tests, more specific relationships for each of the criterion variables are examined by multiple regression analysis. Walberg (1971) suggests a particular order of entry of variables of education studies into multiple regression equations: first, aptitude variables, next, instructional variables, and finally, variables whose role is less well established. Such an order helps to ensure that effects which may be attributed to new variables are significant in themselves and not because of relationships with the well established variables. For the present study the verbal and quantitative ability variables which have a well established role in school performances are treated as covariates in multiple regression equations. Hence the hypotheses concerning the other variables besides verbal and quantitative ability are presented in the order of hypotheses regarding physics achievement followed by hypotheses concerning the cognitive styles.

Hypotheses of Overall Relationships

The set of all variables has two subsets, the criteria and the predictors. A positive relationship is anticipated among the criterion variables and some of the predictors based upon the common physics content of the three criterion variables and the physics achievement predictors. Because of the well-known relationship which exists between school achievement and general ability measures (Cattell and Butcher, 1967), the verbal and quantitative ability variables are expected to be positively related to the variables which possess physics content. The following hypothesis seems justified:

A statistically significant first canonical correlation coefficient exists between the criterion variables and the physics achievement, cognitive style and verbal and quantitative ability variables.

Stated in null form the hypothesis becomes:

1.1 The first canonical correlation is zero between the set of assessment context variables and the set of variables including cognitive styles, intellectual ability and physics achievement.

In addition to the relationship existing between the two sets of variables because of the common physics ability, a second linkage is anticipated because of a cognitive styles dimension. The two tests low in assessment context, the LCT and the DDT, have item features which seem to require, for mastery, capabilities related to field independence and breadth of categorization. The processes of selecting data subsets from items of the LCT or DDT, of making or testing of hypotheses, or of re-examining the items to select alternative subsets of data appear to be similar to those required in overcoming an embedding context on the HFT. The process of examining ranges of values which constitute reasonable estimates of a quantity as required in the CWS resemble the processes used in estimating responses suggested by various data subsets in the LCT and DDT. The probable relationships of breadth of categorization, extent of field independence and the tests which are low in assessment context are discussed in greater detail in the subsection to follow on cognitive style, but a positive association is predicted between the LCT and DDT of the set of criterion variables and the CWS and HFT of the set of predictor variables. Thus the following hypothesis seems warranted:

> A statistically significant canonical correlation is expected between the criterion and predictor variables based on the effects of cognitive styles.

Stated in null form the hypothesis becomes:

1.2 The second canonical correlation is zero between the set of assessment context variables and the set of variables including cognitive style, intellectual ability and physics achievement.

The Criterion Variables

The items of the HCT assess physics content and have been constructed in accordance with the usual standards for multiple choice items. Although they also assess physics content, the items of the LCT and the DDT, however, contain additional attributes. The items of the DDT are diagrammatic and both the DDT and LCT items possess minimal redundancy of essential information, contain irrelevant physics information and suggest two or more physics concepts for their solution. Except for the diagrammatic nature of the DDT, the above characteristics are the ones which make the DDT and the LCT low in assessment context. Hence there is reason to expect that while part of the variance of the student scores on the three criterion variables may be attributed to differences in ability in physics some of the variance in the scores of the LCT and the DDT may be due to other factors. Therefore, after the effects of physics achievement have been removed, the partial correlation coefficients between the HCT and LCT and the HCT and DDT are expected to be equal to zero. The hypotheses are stated as follows:

2.1 The partial correlation coefficient between the HCT and the LCT is zero when the influence of physics achievement is eliminated.

2.2 The partial correlation coefficient between the HCT and the DDT is zero when the influence of physics achievement is eliminated.

Predictors of Criterion

The hypotheses developed below show how the various predictors can be expected to influence each of the criterion variables.

The covariates. The ability measures, the SCAT-V and SCAT-Q were administered to the students in their ninth year of school, approximately three years before the other measures of the study. An indication of the extent to which ability variables can be expected to be predictive of achievement after an interval of three years is suggested by available evidence of predictability over a two-year interval. Three different schools have reported (SCAT Technical Supplement, 1962) correlations of science achievement of college preparatory students at the end of their 10th year of school with the sum of SCAT-V and -Q scores as determined near the beginning of their ninth year. For the three samples of 78, 35 and 25 students the correlation coefficients were 0.77, 0.38 and 0.24 respectively. Undoubtedly the correlations would be lower after the three-year interval, but probably still appreciable.

The presence of the ability variables as covariates permits a more meaningful interpretation of the distinctive contributions of the other predictors in accounting for criterion variance beyond that predicted by the ability variables.

Physics achievement. The three assessment context tests, the HCT, the LCT and the DDT have been designed in order to investigate the extent to which school subject tests in physics may be assessing cognitive styles in addition to achievement. Hypotheses presented below suggest how physics achievement on two subtests of the Physics 30 Examination (P30) may influence performance on the three assessment context tests.

The two subtests of the P30 are made up of the Avital-Shettleworth Knowledge (ASK) items and the Avital-Shettleworth Algorithmic Thinking (ASAT) items. The items of the ASK require recall or exact repetition of the content of the physics course studied by the students. The ASAT items require students to generalize their knowledge of physics by interpreting or applying it in situations not encountered during the learning process.

As described above, scores on the SCAT-V and SCAT-Q are entered first in multiple regression equations to predict each of the assessment context test scores. Entered second are scores of the ASK. The common physics content of the assessment context tests and the ASK items suggests that the, prediction of scores on the HCT, LCT and DDT are expected in each case to be significantly improved by the ASK scores among the predictor variables. Stated in null form, the three hypotheses are:

3.1 There is no significant increase in the squared multiple correlation coefficient (R^2) between the HCT and the ability variables when the ASK scores are added to the set of predictors.

3.2 There is no significant increase in the squared multiple correlation coefficient (R^2) between the LCT and the ability variables when the ASK scores are added to the set of predictors.

3.3 There is no significant increase in the squared multiple correlation coefficient (R^2) between the DDT and the ability variables when the ASK scores are added to the set of predictors.

To be entered next in the multiple regression equations are scores on the ASAT items. According to Avital and Shettleworth (1968) algorithmic thinking requires symbolization of information presented in verbal or diagrammatic form, manipulation of the symbols according to a rule and possibly the formulation of multistep procedures for solving the problem. Although the items of the LCT, HCT and DDT have not been constructed or classified as knowledge or algorithmic thinking in the level of ability required for solving them, algorithmic thinking as defined by Avital and Shettleworth is likely to be required for some of them. The items of the DDT are different from other test items the student has met and require him to reformulate his perceptions in ways probably unlike those practiced in conventional physics instruction. Some of the verbal items of the LCT and the HCT require multistep procedures while others require students to apply knowledge in situations demanding more than recall or recognition. Hence the hypothesis is made that the addition of ASAT scores to the multiple regression equations yield further significant increases in the predictability of achievement on the assessment context tests. In null form the three additional hypotheses are:

3.4 There is no significant increase in the R^2 between the HCT and the ability variables plus physics achievement at the knowledge level when physics achievement at the algorithmic thinking level is added to the set of predictor variables.

3.5 There is no significant increase in the R^2 between the LCT and the ability variables plus physics achievement at the knowledge level when physics achievement at the algorithmic thinking level is added to the set of predictor variables.

3.6 There is no significant increase in the R^2 between the DDT and the ability variables plus physics achievement at the knowledge level when physics achievement at the algorithmic thinking level is added to the set of predictors.

Breadth of categorization. In a study which seems relevant for comparing predicted performance on the HCT and LCT Messick and Kogan (1965) constructed three quantitative aptitude tests with mathematical content. One of the tests had wide numerical spacings among the response alternatives, another had narrowly spaced response alternatives, while the third was free-response in its format. In two examples presented in the report the item with wide spacings had a ratio of largest to smallest alternative of about 5, while the similar ratio for the narrowly spaced item was 1.4. Unfortunately Messick and Kogan did not state the criterion which was used for distinguishing between items. The three tests were administered to 40 undergraduate males along with the Pettigrew (1958) Category Width Scale (CWS). High scores on the CWS indicate a wide breadth of categorization. Besides large positive correlations, $\underline{r} > .56$, among the three quantitative ability tests, there was one statistically significant correlation, $\underline{r} = .57$, involving the CWS and an ability test. This positive correlation of 0.36 occurred with the test having widely spaced alternatives and the items of the CWS with heavy loadings on a factor derived from a centroid analysis of the CWS as performed by Pettigrew (1958). Messick and Kogan suggested that the broad categorizer may apply an "approximation" strategy for reasoning toward the correct response from among the widely spaced alternatives. The broad categorizer can then check his choice by calculation while the narrow categorizer must rely exclusively upon calculation for arriving at his response. The broad categorizer may lose the advantage of the checking procedure when he must discriminate among alternatives by careful computation on tests with narrowly spaced alternatives.

The items of the LCT bear a resemblance to the widely spaced items and those of the HCT to the narrowly spaced alternatives of the Messick and Kogan (1965) measures. The LCT items always have more than one concept or principle among the answers whereas the HCT items do not. For example with a given LCT item the student may be asked to choose from among quantities of energy, momentum, work, velocity, or acceleration, whereas with a given HCT item he may be presented with five different amounts of kinetic energy from which to choose. The LCT and HCT items do not differ in a systematic way in the sizes of the number ranges provided in responses. Because of the similarities of the items of the LCT and the HCT respectively to the widely spaced and narrowly spaced items of the study of Messick and Kogan (1965), and because six of the 10 items on the CWS employed in the present study are the same items with heavy loadings on the factor Messick & Kogan found positively related to performances on the test with wide spacings, a positive relationship is anticipated for the CWS and the LCT. But might the predicted positive relationship of the CWS and the LCT exist because a positive association between physics achievement and breadth of categorization and not because of the widely spaced response alternatives of the LCT? The study of Field and Cropley (1970) suggests a negative answer.

Field and Cropley administered Pettigrew's (1958) Category Width Scale and the General Science Test prepared by the Australian Council for Educational Research to a sample of 178 high school seniors in two rural high schools. The ratio of males to females in the sample was about 7:5.

Correlations between breadth of categorization and science achievement were not statistically significant with $\underline{r}$= -0.01 for males and $\underline{r}$= 0.06 for females. The fact that the subjects of the present investigation are high school seniors and predominantly male means that they are similar to the sample of Field and Cropley. While the content of the HCT, LCT, and DDT is physics and not general science the results of the study by Field and Cropley are probably applicable. Thus no association of breadth of categorization and assessment context is anticipated based upon an association of breadth of categorization and physics achievement. The anticipated positive correlation of the LCT and the CWS may be due to the nature of the LCT items and not the physics content of the LCT.

The arguments presented above in favor of a positive association between performance on the LCT items and breadth of categorization do not seem equally valid with respect to the items of the DDT or the HCT. The DDT items are free response, and Messick and Kogan (1965) did not find a significant association of the CWS and free response items on quantitative tests. The HCT items, although multiple choice in format, do not have the diversity of physics concepts among the response alternatives which the LCT items have. The lack of diversity of physics concepts among the response alternatives suggests that few solution hypotheses are likely to be raised in the student's mind. Therefore the approximation procedure postulated by Messick and Kogan (1965) for deciding upon the best response may not have a chance to operate.

In summary then, breadth of categorization is hypothesized as being positively related to performance on the LCT but unrelated to performance on the HCT and DDT. Since breadth of categorization seems to be independent of verbal and quantitative ability and of physics achievement, breadth of categorization is expected to contribute significantly to the prediction of performance on the LCT beyond what is possible from verbal and quantitative ability and from physics achievement.

The hypotheses are restated below in null form:

4.1 There is no significant increase in the R^2 between the HCT and the ability measures plus physics achievement when scores on the CWS are added to the set of predictors.

4.2 There is no significant increase in the R^2 between the LCT and the ability measures plus physics achievement when scores on the CWS are added to the set of predictors.

4.3 There is no significant increase in the R^2 between the DDT and the ability measures plus physics achievement when scores on the CWS are added to the set of predictors.

Field independence. Witkin, Lewis, Hertzman, Machover, Meissner and Wapner (1959) and Witkin, Dyk, Faterson, Goodenough, and Karp (1962) have claimed that field independence involves the ability to overcome an embedding context. In contrast to field dependent people, field independent persons are better able accurately to maintain a vertical orientation of the body even though the visual field would tend to destroy this orientation, to re-align accurately a lighted rod vertically within a lighted square frame even though no other surrounding visual cues are available and the frame and rod are rotated to different angles, and to locate a simple geometric pattern within a complex array (Witkin et al., 1954).

The items of the DDT require the student to attend to various parts of the diagrams. In order to make and check on hypotheses about the diagrams he must examine the orientation of various parts, distances, relative positions and relative sizes and shapes of parts. Therefore, because of similarity with overcoming embeddedness DDT performance should be predictable to a degree from scores on the HFT by which field independence is assessed in the present study.

Witkin et al. (1962, p. 80) claim that field independence is primarily a perceptual trait, and that it is not related to verbal abilities (1962, p. 203). In arriving at this position regarding verbal abilities Witkin et al, (1962) relied upon the findings of Podell and Phillips (1959). Podell and Phillips required their subjects to rearrange letters in both meaningful and nonsense words in order to form new words and they found little evidence that field independence was related to ability to form new words from either actual words, an ability thought to place a considerable requirement on the overcoming of embeddedness, or nonsense words. However, other studies employing verbal ability measures such as the SCAT-V (Spotts and Mackler, 1967), the WAIS vocabulary scale (Wachtel, 1968) and the SAT-V (Highley, 1970) have found significant positive relationships between field independence and verbal ability. The verbal ability measures emphasize word meaning, comprehension and usage rather than the narrower recognition abilities required by Podell and Phillips (1959). Podell and Phillips (1959, p. 451) state specifically at one point that their tests do not require the level of meaningful organization which characterizes sentences. There are grounds, then, to speculate upon a positive relationship of field independence and verbal ability in some testing formats. Wachtel (1972) has suggested that studies involving field independence should control for intellectual ability, of which verbal ability is an important part, in order that the impact of field independence distinct from verbal ability may be fairly assessed. The present study follows Wachtel's recommendation in treating verbal and numerical ability as covariates of field independence in relation to performance on the various assessment context tests.

Evidence of a connection between field independence and method of instruction in school has been provided in a study by Grieve and Davis (1971). The content of the instruction was the geography of Japan and the pupils were ninth grade boys and girls. Essentially the same lesson materials were used in an expository presentation and in a discovery method over 11 hours of instruction. The HFT was administered to pupils who were divided into field dependent and field independent groups according to whether they were above or below the test median. Pupils were randomly assigned to classes in a 2 x 2 methods by cognitive styles design. In an analysis of scores on a knowledge level achievement test administered at the conclusion of the instruction there were no main effects or interactions. But when scores were reanalyzed only for the 25 percent of pupils who were at the extremes in their HFT scores, there was a significant ($p < .05$) method by cognitive style interaction effect for males and a main effect ($p < .05$) of cognitive styles for females. The more field independent males performed better under the discovery method and the more field independent females scored higher than the less field independent females.

In the present study, rather than selecting students who are at the extremes in extent of field independence, the assessment context tests have been constructed in such a way that only high context and low context items are used and items of middle values of assessment context are omitted. As described in Chapter 3 there are several characteristics upon which assessment context depends. Low and high context items have been defined in such a way that they must differ simultaneously on the selected distinguishing characteristics.

A second test administered to pupils by Grieve and Davis (1971) was a test designed to assess ability to use geographic materials similar to those in the instruction. When these scores were analyzed there was a significant main effect of cognitive style for all males ($p < .05$) and for all females ($p < .01$). Field independent pupils had higher achievement in both samples. There was also a significant cognitive style by method effect ($p < .05$) for males.

While Grieve and Davis (1971) study showed that extent of field independence is related to performance in a school situation it did not attempt to show whether or not the more dominant effect of field independence on the second test as compared with the first one was due to differing amounts of achievement in these two areas or whether differences in the nature of the test items accentuated the cognitive style factor. In the present study pupils will have had the same type of instruction and they will be assessed on three context tests designed to "pick up" on certain cognitive style strengths.

Because of its nature the LCT possesses characteristics which should provide the more field independent subjects with an advantage over the less field independent subjects. Since the stems of the LCT items contain information which could suggest the applicability of two or more physics principles or response alternatives, the student presumably has to go back to the item stem and secure data which will enable him to select the best response alternative. The process of examining each of the response alternatives and then going back to the item stem and selecting data to decide whether or not the alternative is the justifiable one should require the overcoming of the embedding context of the item stem a number of times. Therefore extent of field independence is expected to improve the prediction of LCT scores beyond what is possible from verbal and numerical ability and physics achievement.

The HCT items, in contrast to the LCT items, present only information needed to solve the item and that information is given in a way which directs the student to the recall or use of the relevant physics concept. The continuing need to return to the item stem to select different subsets of data does not exist and hence the problem of overcoming an embedding context in accessing the needed information may not occur. Thus for answering items of the HCT there should be no advantage to the more field independent students.

The items of the DDT are rated low in assessment context and as such appear to provide the more field independent student with an advantage in responding to them as do the items of the LCT. The student responding to the HFT of field independence examines complex diagrams for the presence of one of several simple figures; the student examines the diagrams of the DDT in the search for discrepancies between what is portrayed in the figures and what he would expect to see based upon his knowledge of physics. Shapes, relative sizes and spacing of the figures may or may not be relevant in forming or recalling an expectation for what is presented. The examination of the figures, the forming of expected ideas based upon knowledge of physics and the testing of expectations against given figures seems to require in part the overcoming of the embeddedness of what is shown. Therefore extent of field independence is anticipated to aid significantly in the prediction of performance on the DDT. Furthermore, extent of field independence is expected to improve significantly the prediction of DDT performance when added to verbal and numerical ability variables and to physics achievement.

The hypotheses on the prediction of scores on the three assessment context tests from the scores of the HFT for assessing extent of field independence are summarized below in null form:

5.1 There is no significant increase in the R^2 between the HCT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

5.2 There is no significant increase in the R^2 between the LCT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

5.3 There is no significant increase in the R^2 between the DDT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

Level of Significance

The level of significance accepted for statistical tests of hypotheses of the present study is 0.05. This means that null hypotheses are to be rejected only if the probability of the observed magnitudes arising by chance are five percent or less.

Data Analysis Procedures

Hypotheses have been constructed about the extent to which two cognitive styles and physics achievement are related to performance on three tests which vary in assessment context. The three context tests, the criterion tests, purport to measure ability in physics by means of three different kinds of test items. The predictors are scores from two subtests of physics achievement, the ASK and the ASAT subtests, two cognitive style tests, the CWS and the HFT. Verbal and quantitative ability scores, from the SCAT-V and SCAT-Q are treated as covariates. The present study is not experimental in the sense that there are treatment and control groups, but is correlational. Certain relationships which exist in theory are to be tested. After preliminary exploration of the relationships among the three criterion variables through partial correlations, the principal objective of the calculations will be estimation of the covariances among the criterion and predictor sets, and the predictability of individual criterion test performance from seven independent variables. All measurements have been made on a sample of Alberta high school students enrolled in the Physics 30 course, based upon the Stollberg and Hill textbook, during the spring of 1972.

Canonical correlation is an appropriate procedure to use when a single population is involved and where there are two sets of variables, a criterion set and a predictor set (Cooley and Lohnes, 1971). The use of canonical correlations is to be preferred to reliance upon simple Pearson correlations when several variables are involved. A series of univariate statistical tests is unsatisfactory in a situation where all the variables have been obtained from the same sample because the tests are not independent but are correlated in some unknown manner (Bock and Haggard, 1967). Therefore a multivariate test is desirable which takes into account the correlations among the variables and which has sampling distributions permitting the calculation of probabilities. A significance test, developed by Bartlett (1947), is available for canonical correlations.

Following canonical analysis, specified independent variables are entered into multiple linear regression equations for predicting each of the criterion variables.

Canonical Correlation

Hotelling's (1936) theory of canonical correlation is presented in a comprehensive manner by Anderson (1958). Tatsuoka and Tiedeman (1963) have noted that canonical correlation was not widely used prior to 1963, probably because of heavy computational requirements. Since 1963 however programs for the calculation have become available in most data processing centres.

A canonical correlation, or canonical correlation coefficient, is the correlation between the set of weighted predictors and the set of weighted criteria. The two sets of weights are calculated so as to maximize the correlation between them. Suppose $\bar{z}_{ci}$ represents a vector of normalized scores of criterion variables obtained on subject (i), and $\bar{z}_{pi}$ a vector of normalized scores of predictor variables for the same subject. A set of weights $\bar{a}_1$ and $\bar{b}_1$ are calculated and applied to the original variables thereby producing a component criterion score and predictor score for each individual. The components are:

$$x_i = \bar{a}_1' \, \bar{z}_{ci} \text{ and } y_i = \bar{b}_1' \, \bar{z}_{pi} \, .$$

The two sets of weights are calculated so that the correlation between the components is maximized, that is

$$r_{c1} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i \,{}_{\max}$$

Then r_{c1} is the first canonical correlation between the components. If the weights $\bar{a}_1$ and $\bar{b}_1$ are also calculated such that the means (M) of the weighted variables equal zero, that is

$$M_x = M_y = 0 \, ,$$

and such that the components have variances (v^2) equal to unity, that is

$$v_x^2 = v_y^2 = 1 ,$$

then the components may be called factors. In this case the sets of weights, $\bar{a}$ and $\bar{b}$, are known as factor coefficients and x and y as canonical factors.

As an aid to interpreting the canonical correlation the factor structure ($\bar{s}$) is calculated for the criterion set and for the predictor set. The factor structure is the vector of correlations of the original variables and canonical factors, that is

$$\begin{aligned} \bar{s}_{c1} &= \frac{1}{N} \sum_{i=1}^{N} \bar{z}_{ci} x_i \\ &= \frac{1}{N} \sum_{i=1}^{N} \bar{z}_{ci} (\bar{a}'_z \bar{z}_{ci}) \\ &= \frac{1}{N} \sum_{i=1}^{N} \bar{z}_{ci} \, \bar{z}'_{ci} \, \bar{a}_1 \\ &= R_{cc} \, \bar{a}_1 \end{aligned}$$

where R_{cc} is the matrix of correlations among the criterion variables. In a similar way

$$\bar{s}_{p1} = R_{pp} \, \bar{b}_1 \quad .$$

Another canonical correlation (r_{c2}) may be obtained in the same manner as above but with the additional restriction that the second set of canonical factors must be orthogonal to the first set. The number of canonical correlation coefficients which may be derived is equal to the smaller of the ranks of the predictor and the criterion correlation matrices. Usually this number is the number of variables in the smaller set.

Anderson (1952) has shown that the actual calculations proceed with the solving of the equation

$$(R_{cc}^{-1}R_{cp}R_{pp}^{-1}R_{pc} - r_{c1}^{2}I)\bar{a}_1 = 0$$

under the restriction that

$$\bar{a}_1'R_{cc}\bar{a}_1 = 1$$

Subsequently, $\bar{b}_1 = \frac{1}{r_{c1}}(R_{pp}^{-1}R_{pc}\bar{a}_1)$

Interpretation. Already noted above is the usefulness of calculating the correlations between the variables and the canonical factors for each of the criterion and predictor sets. Such calculations show the relative associations of each variable and each of the possible canonical variates.

Bartlett (1947) has developed a procedure for testing the significance of canonical correlations. For testing departure from a multivariate hypothesis, he has shown that a statistic due to Wilks (1932),

$$\Lambda = \prod_{i=2}^{\min(n_c,\ n_p)} (1 - r_{ci}^2)$$

may be approximated with $(n_c)(n_p)$ degrees of freedom, where (n_c) and (n_p) are the number of variables in the criterion set and predictor set, respectively. Now

$$\chi^2 = -[N-1-.5(n_c+n_p+1)] \log_e \Lambda$$

may be used for the significance, where N is the number of subjects in the sample.

If the null hypothesis for the first canonical correlation is rejected, the second may be tested by

$$\Lambda = \prod_{i=2}^{\min(n_c, n_p)} (1 - r_{ci}^2) \text{ and}$$

$$\chi^2 = - [N-1-.5(n_c+n_p + 1)] \log_e \Lambda ,$$

this time with $(n_c-1)(n_p-1)$ degrees of freedom. Similarly third and further correlations may be tested.

Stewart and Love (1968) and Miller (1969) have provided another means for interpreting canonical correlations, that of redundancy analysis. The redundancy index, calculated separately for each canonical correlation, indicates the fraction of the total variance of one set of variables which is predictable from the canonical variate of the other set. That is, the redundancy index for the criterion set of variables is the fraction of the total variance of the criterion variables which is predicted from the predictor variables; similarly, the redundancy index of the predictor set of variables is the fraction of the total variance of the predictor variables which is predicted from the criterion variables.

The redundancy index for a set of variables is calculated as the product of the ratio of the sum of the common variances of the variables and the variate to the total variance of the variables and the squared canonical correlation coefficient between the two sets of variables. The numerator of the ratio referred to above is equal to the sum of the squares of the correlations between the variables and the canonical factor for the set while the denominator is equal to the number of variables in the set since each one is assumed to be standardized with unit variance. The squared canonical correlation coefficient represents the variance of the criterion and predictor variate which is common or, alternatively, which is predictable in one set of variables from knowledge of the other set.

Suppose that the first canonical correlation between a criterion set of four variables and a predictor set of six variables is 0.90. Furthermore suppose that the sum of the squared correlation coefficients between the variables of the criterion set and the first criterion variate is 3.2 and that the sum of the squared correlation coefficients between the variables of the predictor set and the first predictor variate is 3.0. Now the ratio of the sum of the variances of the criterion variables and the first criterion variate to the total variance of the criterion variables is (3.2 ÷ 4.0). The redundancy index value for the criterion set given the predictor set is

$$\frac{3.2}{4.0} \times (.90)^2 = 0.65$$

and the value of the redundancy index for the predictor set is

$$\frac{3.0}{6.0} \times (.90)^2 = 0.41$$

The value of the redundancy index for the criterion variables suggests that the variance of the criterion set has been predicted reasonably well by the other set of variables; on the other hand, the much smaller value of the redundancy index for the predictor set indicates that the predictors are rather inefficient in the sense that a relatively small fraction of predictor variance was effective along the first variate.

In the general calculation of the redundancy index for the criterion set, the variance of the criterion variables and the first criterion variate is the sum of squares of the factor structure coefficients, s_c , that is

$$\bar{s}'_{1c}\ \bar{s}_{1c}$$

The ratio of the variance of the criterion variables and the first canonical variate to the total variance of the criterion set is:

$$\frac{\bar{s}'_{1c}\ \bar{s}_{1c}}{n_c}$$

where n_c is the number of criterion variable. The redundancy index ($\underline{r}_d$) for the criterion set on the first variate is

$$r_{dc1} = \frac{\bar{s}'_{1c}\ \bar{s}_{1c}}{n_c} \cdot r_{c1}^{2}$$

Similarly, the redundancy of the predictor set on the first variate is

$$r_{dp1} = \frac{\bar{s}'_{1p} \cdot \bar{s}_{1c}}{n_p} \cdot r_{c1}^{2}$$

The concept of the redundancy index may be extended to the total redundancy of the criterion set for all variates

$$r_{dp} = \sum_{i=1}^{\min(n_c,n_p)} \frac{\bar{s}'_{ic} \; \bar{s}_{ic}}{n_c} \cdot r_{ci}^{2}$$

and for the predictor set for all variates

$$r_{dp} = \sum_{i=1}^{\min(n_c,n_p)} \frac{\bar{s}'_{ip} \; \bar{s}_{ip}}{n_p} \cdot r_{ci}^{2}$$

Multiple Regression Analysis

Multiple linear regression analysis is similar to canonical analysis except that there is but one criterion variable instead of a set of several. Therefore, the equation to be solved for canonical correlation

$$\left[R_{cc}^{-1} \; R_{cp} \; R_{pp}^{-1} \; R_{pc} - r_c^{2} I\right]\bar{a} = 0$$

is considerably simplified in the multiple linear regression situation, since with only one criterion R_{cc} and R_{cc}^{-1} are unity, R_{cp} is $\bar{r}_{1p}$, a= 1 and the equation is no longer a matrix equation. As shown by Cooley and Lohnes (1971) the vector of weights of the predictors, $\bar{b}$, which minimizes the prediction error in the least least squares sense is

$$\bar{b} = R_{pp}^{-1} \; \bar{r}_{1p}$$

and the squared multiple correlation coefficient, R^2, is

$$R^2 = \bar{b}' \bar{r}_{p1}$$

From the above equations it can be seen that if the predictors are perfectly non-correlated, the matrix R_{pp} and its inverse R_{pp}^{-1} equal the identity matrix; the $\bar{b}$ and R^2 would depend exclusively upon the intercorrelations of the criterion and the various predictors, that is upon $\bar{r}_{1p}$. Furthermore, the dropping of one predictor makes no difference to the remaining weights in $\bar{b}$, but the size of R^2 will change.

If the predictors are positively correlated as some in the present study probably will be, Darlington (1968) has described how an error of overestimation in one weight will be compensated for by an underestimation of the other weight. The error in overestimation in one weight for the population could occur because of lack of randomization of the sample. Darlington (1968) has stated that the larger R, the greater is this compensatory effect. Thus it is not uncommon with highly correlated predictors to find that two regression equations from two different samples of the same population may have widely different weights but about equally good prediction. The implication of this principle is that with highly intercorrelated predictors, interpretation in terms of R^2 is more reliable than in terms of $\bar{b}$.

The statistical significance of R^2 may be evaluated by means of an F-test. McNemar (1969) has demonstrated that with m predictor variables and N subjects the test is

$$F = \frac{R^2/m}{(1 - R^2)/(N - m - 1)}$$

where there are m degrees of freedom for the numerator and (N - m - 1) for the denominator.

Multiple linear regression hypotheses of the present study are stated in terms of the improvement in prediction which results when specified predictors are added to those already included in the regression equation. If the variance of the criterion accounted for by m predictors is R_m^2, and the variance accounted for when an additional n predictors are added is R_{m+n}^2, the improvement in prediction may be tested by

$$F = \frac{(R_{n+m}^2 - R_m^2)/n}{(1 - R_{n+m}^2)/(N - m - n - 1)}$$

where there are n degrees of freedom for the numerator and (N - m - n - 1) for the denominator.

Cohen (1965, 1968) has shown that regression analysis is robust with respect to violations of assumptions of normality and homogeneity of variance of the dependent variable for any given combination of predictors.

When making hypotheses which are to be tested by multiple linear regression, Cohen (1968) has pointed out that covariates should enter the equation first, followed by the variables likely to be of most relevance. Then a second set of variables should enter consisting of lower order interactions and possibly some quadratics with the stipulation that these variables are to be viewed less as hypotheses and more as exploratory issues. Walberg (1971) has speculated that in education studies aptitudes should be entered before instructional and environmental variables. In the present study verbal and numerical abilities scores are covariates and hence are entered first in the prediction equation. Placed next is achievement in physics, an instructional variable. The reasons why the cognitive style variables have been entered after physics achievement have been elucidated earlier in this chapter. Exploratory variables with interactions and quadratic terms will be entered last.

CHAPTER 5

RESULTS AND DISCUSSION

The chapter presents the results of the study and a discussion of their meaning beginning with an exposition of the properties of the various tests which were administered. Included is an outline of the procedures used in dealing with missing data and a discussion of the matrix of intercorrelations among the measures. Presented next are the results of tests of the various hypotheses. The multivariate hypotheses are tested first followed by hypotheses of relationships among the criterion variables and finally of hypotheses concerning the prediction of each of the criterion variables in turn.

All major calculations have been performed with the aid of computer programs available through the Division of Educational Research Services, The University of Alberta.

Test Results

This section begins with the results of the rating of items according to the levels of the Avital and Shettleworth (1968) categories. The extent of missing data is presented next, followed by test results and intercorrelations. The test scores of the students, with test means substituted for missing data points, are presented in appendix D.

Preliminary results of physics achievement item ratings. The items of the Physics 30 Test (P30) were rated as to membership in the Avital and Shettleworth (1968) categories. The rating yielded only eight items in the knowledge category and therefore the Avital-Shettleworth Knowledge subtest (ASK) was a test of only eight items. Such a short test has very doubtful content validity since it does not sample adequately from the concepts of a full year physics course; furthermore the reliability is probably inadequate. The rating of P30 items yielded 48 at the algorithmic thinking level, eight at the higher taxonomic levels and one item which could not be rated as it did not seem to assess a concept which was included in the Physics 30 Course. The decision was made to abandon the plan to utilize two subscores, the ASK and the ASAT, from the P30. Instead the P30 was retained intact as a unitary measure of physics achievement, except for the one item judged invalid. Scores on this item were not included in the P30 results.

That only eight items of the P30 were rated as being in the higher taxonomic levels, that is, in the Open Search category of Avital and Shettleworth (1968), is not surprising. Avital and Shettleworth (1968, p. 43) note that open search problems sometimes require long periods of deliberation and the controlled conditions of classroom testing do not readily lend themselves to the making of assessments of open search. More fundamentally, Avital and Shettleworth (1968, p. 36) have suggested that the ability to solve open search problems does not follow from instruction as closely as do abilities in the knowledge and algorithmic thinking areas. Item writers may feel that open search items are not valid in the sense that they do not assess what has been taught and therefore tend to reject such items. A combination of this concern about instructional validity and the unpredictable time requirement for solving some open search problems are likely the reasons behind the small number of such items in the P30.

Missing data. The present study was planned so that measurements would be made on nine variables. However, as described above, the P30 has been used as a single measure of physics achievement rather than two subscores, the ASK and the ASAT. Therefore, measurements are reported on eight variables.

The data for any member of the sample who was missing tests scores for more than two of the criterion test scores, or for more than three of the eight tests were eliminated from all calculations. Test results of five persons, three male and two female were removed as a result of the missing data decision. The size of the sample was thereby reduced to 139, with 119 males and 20 females.

Numerical values are provided in Table 5 on the extent of missing data for each of the tests. The HCT, LCT and DDT are the criterion tests and the SCAT-V, SCAT-Q, PT, CWS and HFT the predictors.

The tests, except for the SCAT-V and SCAT-Q, had to be

TABLE 5

MISSING DATA IN THE SAMPLE

OF 139 SUBJECTS

Test	Number of missing points
HCT	7
LCT	7
DDT	19
SCAT-V	29
SCAT-Q	29
P30	0
HFT	19
CWS	23

administered near the end of the school term when the teaching of the Physics 30 course was nearly completed as the HCT, LCT, DDT sampled from concepts throughout the course. Attendance near the end of the school term was poorer than average which is the reason for most of the missing data. The SCAT-V and SCAT-Q were administered as part of a province-wide testing program of ninth grade students. Since the subjects of the present sample were twelfth grade pupils many of the missing SCAT scores are accounted for by students who have moved to Alberta during the intervening period.

Following the advice of Cohen (1968), missing data points for persons not eliminated from the sample may be replaced by the means of the tests for which data are missing. Cohen (1968) has noted that this substitution of the means recognizes that the population itself has missing data; furthermore, the substitution of test means is a conservative procedure since the degrees of freedom for testing statistical significance are increased even though the magnitude of the correlation is not altered by the substitution of a test mean for a missing data point.

No substitutions of test means for missing data were made before test means, standard deviations and reliabilities had been calculated as presented in Table 6. But all results subsequent to the initial calculations have had substitution for missing data.

Means, standard deviations and reliabilities. Reported in Table 6 are the means, standard deviations, maximum possible scores, number of subjects tested in the present study, and reliabilities for the tests. The scores from the five subjects who exceeded the missing data criterion were eliminated prior to the calculations.

The scores for each test were encompassed within eight equal-sized intervals and tested against a normal distribution by the calculation of a *chi-square* value. Except for the HCT all of the *chi-squares* were small enough, with $p > .07$, to suggest that the assumption of normality was warranted. The HCT scores were positively skewed and there was a probability of only 0.3 percent that a greater deviation from normality would occur by chance.

The reliabilities presented in Table 5 are K-R 20 measures of internal consistency. The speeded nature of the HFT may have resulted in a spuriously high value for the reliability. Test-retest reliabilities for the HFT have been reported at 0.63 over a 10-week interval (Boersma, 1968) and at 0.72 with test administrations separated by only a few hours (Fleishman and Dusek, 1971).

The assessment context tests were difficult tests with the HCT and the DDT the most difficult of them. The average difficulty of the items of the HCT and the DDT was approximately 0.37 which probably resulted in an internal consistency reliability lower than would have been the case if the average difficulty were nearer 0.50.

TABLE 6

SUMMARY OF TEST RESULTS

Test	Maximum possible score	N	Mean	SD	Reliability
HCT	20	132	7.4	2.7	0.49
LCT	20	132	9.5	3.1	0.56
DDT	20	120	7.5	2.9	0.61
SCAT-V	60	110	44.1	9.8	0.90[1]
SCAT-Q	50	110	35.8	7.0	0.90[1]
P30	64	139	36.7	11.4	0.89
HFT	32	120	10.9	5.5	0.79[2]
CWS	60	116	29.4	9.4	0.83

[1] As noted by Green (1965)

[2] As reported by Boersma (1968)

Intercorrelations. The correlations among the tests are presented in Table 7. As expected there are statistically significant positive correlations among many of the variables, especially the assessment contexts tests and the P30. These large correlation coefficients are to be expected because of the common physics content of the assessment context tests.

The CWS is unrelated to the other tests, except for SCAT-V and SCAT-Q. The positive association of the CWS and SCAT-Q, with $r = 0.20$ and $p < .05$, is similar to Pettigrew's (1958) finding of $r = 0.26$ between the category width scale and the American Council on Education test of quantitative ability. It should be remembered that the CWS includes only 10 of the original 20 items of the scale developed and used by Pettigrew (1958). A coefficient $r = 0.36$, $p < .05$ was reported by Messick and Kogan (1965) between one factor of Pettigrew's (1958) scale and an arithmetic ability test with widely spaced response alternatives; four items of the CWS coincided with heavily weighted items employed by Messick and Kogan (1965). Hence the finding of the present study supports the results of Pettigrew (1958) and Messick and Kogan (1965) and strengthens the concept of a moderate association of breadth of categorization and quantitative ability, at least for college and near-college students.

The significant relationship found between the CWS and SCAT-V, $r = 0.22$, $p < .01$, is not in agreement with previous

TABLE 7

MATRIX OF INTERCORRELATIONS[1] OF TEST SCORES

N=139	HCT	LCT	DDT	SCAT-V	SCAT-Q	P30	HFT	CWS
HCT								
LCT	54**							
DDT	55**	49**						
SCAT-V	15	24**	11					
SCAT-Q	20*	17*	15	54**				
P30	63**	62**	59**	36**	41**			
HFT	34**	25**	19*	14	07	29**		
CWS	11	05	12	22**	20*	10	07	

[1] Decimals have been omitted

* $p < .05$

**$p < .01$

findings. The original Pettigrew (1958) CWS was found to be unrelated to verbal ability as assessed by the SAT-V (Kogan and Wallach, 1964) or to verbal IQ as measured by the American College Test (Jackson and Pedersen, 1965).

The HFT is positively correlated with all measures but significantly so only for the tests with physics content; $p < .05$ with the DDT and $p < .01$ for the HCT, LCT and PT. The non-significant correlation coefficients between the HFT and the ability measures, $r = 0.14$ for SCAT-V and $r = 0.07$ for SCAT-Q, are not in agreement with the findings of Dubois and Cohen (1970) but are in agreement with those of Fleishman and Dusek (1971). The significantly positive correlations between the HFT and all of the measures of physics achievement indicate that the more field independent students tend to be the higher achievers in physics, a result not in agreement with the finding of Brilhart and Brilhart (1971). They administered the HFT to a sample of 184 males at the beginning of a college course in engineering, and traced the relationship between HFT performance and achievement in engineering over a 9-term period. No association of field independence was detected. However, the sample of Brilhart and Brilhart (1971) seemed to have been an unusually field independent group. Brilhart and Brilhart (1971) noted that the students were given only 10 minutes to attempt 16 items on the HFT while the usual administration procedure, followed in the present study, allows two 10-minute periods to attempt 32 items. In spite of the abbreviated administration time, Brilhart and Brilhart (1971) reported that 60 per cent of their subjects scored 13 or higher on the 16-item test, whereas in the present study the mean score was 10.9 for a test twice as long. Apparently the engineering students were a highly field independent group and the clustering of HFT scores at the upper end of the scale may not have permitted the full range of field independence tc emerge. The higher physics achievement of the more field independent students in the present study can be interpreted in the light of Witkin's (1972) observation that college entrance students who are highly field independent had enrolled in more optional high school mathematics and science courses than had the less field independent students. The pattern of choice of optional subjects may be due in part to the greater success in science and mathematics of the highly field independent subjects. Bowles and Boss (1974) showed that highly field independent, male students exhibited significantly higher science achievement than the less field independent students even at the grade nine level.

The correlation coefficients between the SCAT-V, the LCT, the P30 and the CWS were all significantly positive, with $p < .01$. The well-established relationship of verbal ability and science achievement (Cattell and Butcher, 1968, chap. 3) is apparent for SCAT-V and the P30. The lack of a significant association between the SCAT-V and the HCT is surprising, since the HCT is a verbal test and has been constructed according to the accepted criteria for a good classroom test.

The SCAT-Q was found to be positively correlated with the three assessment context tests although the values are significant, $p < .05$, only for the HCT and LCT.

As expected, significant positive relationships, $p < .01$, were found to exist among all the measures of physics achievement.

Tests of Multivariate Hypotheses

A significant first canonical correlation was predicted between the set of criterion variables and the set of predictor variables based mainly upon the physics content of the assessment context tests of the criterion set and the physics achievement test and the ability measures of the predictor set. The following hypothesis was presented:

1.1 The first canonical correlation between the set of assessment context variables and the set of variables including cognitive style, intellectual ability and physics achievement is zero.

The next hypotheses predicted a second significant canonical correlation with important contributions to the criterion variate from the low assessment context tests and to the predictor variate from the cognitive style variables. The following hypothesis was presented:

1.2 The second canonical correlation between the set of assessment context variables and the set of variables including cognitive style, intellectual ability and physics achievement is zero.

The canonical correlations and the results of significance tests are presented in Table 8. The first canonical correlations coefficient, r= 0.76, is significantly different from zero, $\chi^2(15)= 124$, $p < .001$. Therefore, hypothesis 1.1 is rejected. The second canonical correlation coefficient, r= 0.19, is not significantly different from zero, $\chi^2(8)= 8.0$, $p = .30$. Hence, hypothesis 1.2 is not rejected.

The rejection of hypothesis 1.1 seems to confirm the predicted relationship between the two sets of variables. Interpretation of the relationship is possible from the structure coefficients which are the correlations of the original variables and the canonical variates, and from the values of the redundancy indices (Stewart and Love, 1968). The structure coefficients and the redundancy indices for the first and second canonical correlations are presented in Table 9.

The vectors of weights which were applied to the criterion and predictor variables in the calculation of each canonical correlation are less reliable for interpreting canonical correlations than the structure coefficients (Cooley and Lohnes, 1971, chap. 4; Meredith, 1964). The vectors of weights to be applied to the original variables, or canonical vectors, are presented in appendix E.

The statistically significant first canonical correla-

TABLE 8

Canonical Correlation Coefficients and Bartlett's Significance Test Results

Canonical Correlation	r	Λ	χ^2	df	p
First	0.75	0.43	124.4	15	$<.001$
Second	0.19	0.94	8.0	8	$>.05$
Third	0.16	0.97	3.3	3	$>.05$

TABLE 9

Structure Coefficients For Criterion and Predictor Variables For the First and Second Canonical Correlations

Criterion Variable		Predictor Variable	
First Canonical Correlation			
High Context Test	.87	SCAT-Verbal	.27
Low Context Test	.87	SCAT-Quantitative	.27
Discrepancy Detection Test	.80	Physics 30 Test	.98
		Hidden Figures Test	.42
		Category Width Scale	.14
Variance Fraction	.68	Variance Fraction	.26
Redundancy Index	.40	Redundancy Index	.15
Second Canonical Correlation			
High Context Test	.39	SCAT-Verbal	.61
Low Context Test	.55	SCAT-Quantitative	.15
Discrepancy Detection Test	.13	Physics 30 Test	.05
		Hidden Figures Test	.33
		Category Width Scale	.35
Variance Fraction	.16	Variance Fraction	.14
Redundancy Index	.01	Redundancy Index	.01

tion means that there is a factor, or canonical variate, in the space of the criterion variables which is highly correlated with a canonical variate in the space of the predictor variables. The structure coefficients for each set of variables consists of the Pearson correlations between the original variables and the respective variate. For the criterion set, the structure coefficients between the first criterion variate and the HCT, LCT and DDT were, respectively, 0.87, 0.82, 0.80. The large and similar values of the three correlations indicate that the three assessment context tests are closely related to the variate. Inspection of the structure coefficients between the predictor variables reveals that the correlation for the P30 test is remarkably large, 0.98, and much greater than any of the other coefficients. Hence, physics achievement is closely related to the first predictor variate. The overall picture which emerges is one with the first predictor variate closely associated with the physics achievement test, with the first criterion variate strongly associated with all three assessment context tests, and, with the criterion and predictor variates themselves significantly related as shown by the first canonical correlation coefficient, $\underline{r}= 0.76$, $\chi^2(15)= 124.4$, $\underline{p} < .001$. The physics content of three assessment context tests and the P30 is probably responsible for the significant overall relationship as had been predicted. The structure coefficient between the first predictor variate and the HFT is 0.42. While the value of the coefficient is much smaller than that for the P30, the HFT bears a reasonably healthy relationship to the predictor variate which is closely related to physics achievement. Such a relationship was anticipated from the significant associations of the HFT and the HCT, LCT, DDT, and P30 on univariate tests, but the canonical analysis has shown that the relationship holds for a predictor variate which correlates significantly with a criterion variate which is a factor of the assessment context tests. Therefore, the association of extent of field independence and physics achievement is quite general within the limits of the present study.

Calculations leading to the redundancy index for the criterion set, given the availability of the predictors, show that the variance ratio of the first criterion variate is 68 percent of the summed variance of the three criteria. The redundancy index for the criterion set, the product of the variance fraction and the square of the first canonical correlation, is 0.40. This means that of the 68 percent of criterion variance which is contained in the first criterion variate, 58 percent of it is accounted for by the predictors; that is, 40 percent of the total variance of the criterion variables is accounted for by the predictors. Similar calculations for the predictor set reveal that 26 percent of predictor variance is utilized in the first predictor variate, and since 58 percent of the variance of the predictor variate is involved in the canonical correlation, only 15 percent of the summed variance of the predictor variables is accounted for by the presence of the three criterion variables.

A comparison of the variance fractions for the predictor and criterion sets shows that 68 percent of available criterion variance is embodied in the first criterion variate, but only 26 percent of predictor variance is in the first predictor variate. At this stage, the possibility exists of a second significant canonical correlation. The second pair of variates which are orthogonal to the first could conceivably utilize the 32 percent of the variance of the criterion variables and the 74 percent of the variance of the predictor variables, variances which were not utilized in the first pair of canonical variates.

The second canonical correlation coefficient, 0.19, is however non-significant. Furthermore, the fractions of variance of predictor and criterion variables utilized in the second set of variates as shown in Table 9 are quite small. Hence the negligible redundancy indices for both predictor and criterion sets are not surprising. This means that canonical variates could not be found which were orthogonal to the first variates and which were highly correlated. Thus the hypothesized relationship between the tests low in assessment context and the cognitive style variables has not been shown to exist.

Since each successive canonical correlation coefficient is smaller than the previous one, and since the second canonical correlation has not reached significance and has utilized only a small fraction of the unaccounted for variance of the predictor and criterion variables, the structure coefficients and redundancy indices for the second and third canonical correlations will be small and non-significant and have not been calculated.

Relationships Among Criterion Variables

In order to assess the impact of the assessment context features of the criterion tests two hypotheses were constructed concerning the relationships among the criterion tests. The hypotheses predicted that the partial correlations between the HCT and the LCT, and between the HCT and the DDT would each become zero when the influence of physics achievement is removed. As stated in the usual form:

2.1 The partial correlation between the High Context Test (HCT) and the Low Context Test (LCT) is zero when the influence of physics achievement is eliminated.

2.2 The partial correlation between the HCT and the Discrepancy Detection Test (DDT) is zero when the effects of physics achievement are removed.

The partial correlation coefficient between the HCT and the LCT was calculated with the influence of the P30 removed. The resulting partial correlation coefficient was

0.24, $t = 2.94$, p (two-tailed) $< .01$. Hypothesis 2.1 has therefore not been accepted. Thus an important residual relationship exists between the HCT and the LCT after the common influence of the PT has been removed. The unique features of the low assessment context items of the LCT had been expected to have a variance component associated with them which was unique and not shared with the variance of the HCT after the variance associated with physics achievement had been removed. The correlation coefficient between the HCT and the LCT was calculated to be 0.54 indicating a common variance of 0.29, and the partial correlation coefficient between the HCT and the LCT with the influence of the PT removed was 0.24 indicating a common variance of 0.06. Therefore the proportion of overlap resulting from the effects of the PT is 0.29 - 0.06= 0.23. The percentage of the total association present resulting from physics achievement is (0.23/0.29) (100)= 79 percent. The remaining common variance, 21 percent, may have been due to elements present in the coinciding administration of the HCT and the LCT which consisted of subsets of items within the test known as the Physics Test. The joint administration of the two subtests tends to capitalize on the effects of situational factors present (Messick and Kogan, 1965) and thus the 21 percent of common variance after the effects of physics achievement were removed may be accounted for at least in part.

For testing hypothesis 2.2 the partial correlation coefficient was calculated between the HCT and the DDT with the effects of physics achievement on the P30 removed. The resulting coefficient equalled 0.28, $t = 3.49$, p(two tailed) $<$ .01. The anticipated null result was not found and hypothesis 2.2 not accepted. As in the relationship of the HCT and the LCT, there was an appreciable and unanticipated common variance for the HCT and the DDT which remained after the influence of the P30 was removed. The correlation of the HCT and DDT was 0.56 indicating a common variance of $(0.56)^2 = 0.31$ and the partial correlation of 0.28 means that variance of $(0.28)^2 = 0.08$ remained in common after the influence of the P30 was removed. Thus $(0.31 - 0.08) = 0.23$ of the variance was $(0.08 \div 0.31 \cdot (100) = 27$ percent of the original common variance.

The appreciable remaining variances of the HCT and the DDT after the influence of physics achievement was removed are difficult to explain. The HCT is a verbal, multiple choice test while the DDT is a diagrammatic, mostly non-verbal free response test. The HCT consists of items which are claimed to be high in assessment context with the DDT items low in assessment context. The HCT and the DDT were administered as separate tests with a week between administrations, not combined in a single test as were the HCT and the LCT.

Although not required for purposes of hypothesis testing, the partial correlation of the LCT and the DDT upon removal of the influence of the PT was calculated and was found to be 0.19, $t = 2.32$, p(two-tailed) $< .02$. The magnitude of the significant positive correlation could be explained on the basis of the low assessment context of the items of the LCT and the DDT; however the magnitude of the partial correlation coefficient between the LCT and the DDT is comparable with those for the HCT and LCT or the HCT and the DDT which did not share in the low assessment context of the items. In summary, all three of the assessment context tests are significantly positively correlated even after the influence of the common physics content is removed.

Multiple Linear Regression Tests of Hypotheses

All of the multiple linear regression equations constructed in the present section have been used to assess hypotheses on the influence of specific variables in predicting each of the criterion variables. In a later section the results of exploratory analyses are described to indicate the presence of unanticipated effects which may have been present.

The hypotheses to be investigated have specified a predetermined order for the addition of the predictors. Therefore, the more common practice of step-wise regression in which the variable correlating most highly with the criterion is entered first in the regression equation has not been followed. Instead a computer program has been used which permits the specification of the order of entry of the predictors. In each of the subsections below the results of tests on the effectiveness of various independent variables in multiple linear regression equations for predicting, in turn, the HCT, the LCT and the DDT are presented.

Verbal and quantitative ability. In each of the multiple linear regression equations scores on the SCAT-V and SCAT-Q have been entered first in the equations and all hypotheses involve the prediction of variance in the assessment context variables beyond what is possible from knowledge of verbal and quantitative ability.

Tests with the HCT as criterion. Hypotheses have been constructed concerning the variance of the HCT predictable from the measures used to assess, in turn, physics achievement, breadth of categorization, and extent of field independence.

Separate hypotheses were developed stating the influence of physics achievement at the knowledge level and physics achievement at the algorithmic thinking level. However, when the items of the P30 were assessed as to their Avital-Shettleworth (1968) categorizations, insufficient items were found at the knowledge level to warrant the use of separate subscores for physics achievement at the knowledge level. Therefore the hypotheses concerning the physics achievement at the knowledge level and the algorithmic thinking levels have been combined for predicting the effect on the HCT of physics achievement. Thus the two hypotheses 3.1 and 3.4 have been combined as hypothesis 3.1 - .4. The original hypotheses and the newly-formed hypothesis are presented below.

3.1 There is no significant increase in the squared multiple correlation coefficient (R^2) between the HCT and the ability variables when physics achievement at the knowledge level is added to the set of predictor variables.

3.4. There is no significant increase in the R^2 between the HCT and the ability variables plus physics achievement at the knowledge level when physics achievement at the algorithmic thinking level is added to the set of predictor variables.

3.1-.4 There is no significant increase in the R^2 between the HCT and the ability variables when physics achievement is added to the set of predictor variables.

The predicted effect of physics achievement for predicting the variance of the HCT was a significantly positive one. The field independence measure, the HFT and the breadth of categorization test, the CWS, were not anticipated as being significantly predictive of the variance of the HCT. The precise statements of the null effects of the HFT and the CWS are presented as follows.

4.1 There is no significant increase in the R^2 between the HCT and the ability measures plus physics achievement when breadth of categorization is added to the set of predictor variables.

5.1 There is no significant increase in the R^2 between the HCT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

The results of the sequential tests of hypotheses are presented in Table 10. At each stage the predictor was entered into the multiple linear regression (MLR) equation and its increment in the squared multiple correlation (R^2) coefficient was tested. If statistically significant, the variable was retained in the equation for succeeding steps; if not, the variable was dropped prior to the next step.

Hypothesis 3.1-.4 was not accepted as the increase in R^2 upon the addition of the P30 variable was 0.37 with $\underline{F}(1,135) = 85.1$, $\underline{p} < .01$. The anticipated influence of the P30 in predicting scores on the HCT was readily observed. The physics content of the P30 and of the HCT was in all probability the reason for the high degree of effectiveness of the P30 in improving the prediction of the HCT.

Tested next was hypothesis 4.1 concerning the effectiveness of the CWS as a predictor of the HCT. The null hypothesis was not rejected since the increase in R^2 upon the addition of the CWS was 0.01, $\underline{F}(1,135) = 0.9$, $\underline{p} > .05$. As expected, the CWS did not improve the prediction of the

Table 10

Tests of Hypotheses on Changes in the Multiple Correlation of the HCT as the Criterion and Various Predictors

Step	Variable(s) Added	R^2	ΔR^2	F	df	p	Decision on Variable
0	SCAT-V and SCAT-Q	.04	.04	2.88	2,136	.05	Covariates
1	P30	.41	.37	85.1	1,135	.01	Add
2	CWS	.42	.01	0.9	1,134	.32	Drop
3	HFT	.44	.03	6.8	1,134	.02	Add

Regression equation at $p < .05$:

$$X_{HCT} = - 0.09\,X_{SCAT\text{-}V} - 0.02X_{SCAT\text{-}Q} + 0.63\,X_{P30} + 0.18X_{HFT}$$

scores on the HCT beyond what was possible from the ability measures and the physics achievement test. The result is in agreement with a previous report that science achievement and breadth of categorization are not related for senior high school students (Field and Cropley, 1970). Although a positive association of breadth of categorization and quantitative ability has been suggested (Pettigrew, 1958; Messick and Kogan, 1965) the quantitative content of the HCT items may not have been important enough, or else the nature of the single concept response alternatives may have been perceived as sufficiently narrowly spaced that the CWS did not improve prediction of the HCT.

The next hypotheses tested concerned the effect of adding the field independence variable to the set of predictors. For the test high in assessment context extent of field independence was not expected to improve prediction, and the null result was stated in hypotheses 5.1. The calculations indicated an increase in the squared multiple correlation coefficient of 0.03, $\underline{F}(1,134) = 6.8$, $\underline{p} < .02$. Therefore hypothesis 5.1 was rejected. Evidently field independence makes a contribution toward performance on the physics test high in assessment context with the contribution independent of physics achievement. The more field independent persons tend toward higher achievement on the HCT. The tendency which has been noted (Witkin, 1972) for the more field independent high school students to select more optional science and mathematics courses may be due to the mutual reinforcement of physics achievement and field independence which seems to be indicated by the present result. Whether the reinforcement occurs within the method of measurement of physics achievement or more broadly in the teaching and learning of the subject field may be indicated by the tests of hypotheses concerning field independence and the LCT and the DDT.

In summary, the prediction of the HCT criterion was significantly improved by the addition of physics achievement scores to the ability variables, and a further significant improvement in prediction occurred with the addition of the field independence measure, but not with the breadth of categorization test.

Tests with the LCT as criterion. The hypotheses to be tested relate to physics achievement, breadth of categorization and extent of field independence. The original hypotheses 3.2 and 3.5 concerning physics achievement at the knowledge and algorithmic thinking levels were combined because of the non-availability of separate scores for the two levels. The combined hypothesis is presented below.

3.2-.5 There is no significant increase in the R^2 between the LCT and the ability variables when physics achievement is added to the set of predictors.

The hypotheses pertaining to breadth of categorization and extent of field independence follow:

4.2 There is no significant increase in the R^2 between the LCT and the ability measures plus physics achievement when breadth of categorization is added to the set of predictor variables.

5.2 There is no significant increase in the R^2 between the LCT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

The results of the MLR tests are presented in Table 11 and are discussed below.

The variance overlap of the ability variables and the LCT is 0.06, an amount which is unimpressive in magnitude considering that verbal and numerical ability are usually good predictors of success in academic school subjects. In the present study the time lag of nearly three years between the administration of the SCAT and the other measures is probably responsible for the weak association between the LCT and the SCAT. Several studies of the prediction of school grades in science from SCAT scores obtained two years previously have been reported in the SCAT Technical Manual (1962, p. 11). While the sizes of samples of high school college preparatory students were all 78 or less, the reported common variances of science achievement and the

Table 11

Tests of Hypotheses on Changes in the Multiple Correlation of the LCT as the Criterion and Various Predictors

Step	Variable(s) Added	R^2	ΔR^2	F	df	p	Decision on Variable
0	SCAT-V and SCAT-Q	.06	.06	4.2	1,136	.05	Covariates
1	P30	.39	.33	75.5	1,135	.01	Add
2	CWS	.39	0	0	1,134	.90	Drop
3	HFT	.40	.01	1.0	1,134	.33	Drop

Regression equation at $p < .05$

$$X_{LCT} = 0.08X_{SCAT-V} - 0.14X_{SCAT-Q} + 0.65X_{P30}$$

SCAT ranged from 0.06 to 0.59 with a median of 0.15. Hence the magnitude of common variance found in the present study is not unrealistically low for a three-year period.

As expected, the P30 test of physics achievement produced a statistically significant improvement in the prediction of the LCT, undoubtedly because of the common physics content, $\Delta \underline{R}^2 = .33$, $\underline{F}(1,135) = 75.5$, $\underline{p} < .01$.

Null hypothesis 4.2, that the CWS does not improve the prediction of the HCT, was tested next and was not rejected, $\Delta \underline{R}^2 = 0$, $\underline{F}(1,135) = 0$, $\underline{p} = .90$. The result means that, contrary to expectations, knowledge of breadth of categorization scores did not improve the prediction of performance on the LCT beyond what was possible from ability measures and physics achievement. Apparently the approximation strategy for deciding among conceptually wide-spaced response alternatives did not operate with the low assessment context items of the LCT as Messick and Kogan (1965) suggested it might for deciding among the responses to items with widely spaced alternatives. Various explanations are possible of the failure to substantiate the predicted relationship. The approximation strategy of Messick and Kogan (1965) may not be correct. They observed it only for one factor of Pettigrew's (1958) original CWS, and the factor for which they observed the significant relationship with the quantitative test of widely spaced alternatives was not the factor for which Pettigrew (1958) had reported a significant correlation with the American Council on Education test of quantitative ability.

If Messick and Kogan's (1965) hypothesis is correct, the relationship between the CWS and the LCT may not have emerged since only four items of the 10-item CWS used in the present study had weights greater than 0.50 in determining the factor found to be significantly correlated with the test having wide-spaced alternatives. Still another possibility is that the items of the LCT, with 12 of 20 items not having numerical quantities among the response alternatives, were not sufficiently similar to the items of Messick and Kogan's (1965) study for the same kind of approximation strategy operates mainly with items having numerical responses. There is another reason to expect that the approximation strategy should have been forced on the LCT. The average difficulty of the items on the LCT was 0.48, and the difficult nature of the test might have resulted in more than average levels of guessing at answers, and guessing is an approximating strategy. Whatever the reason, the process of estimating the farthest value from the mean as required in the assessment of breadth of categorization by the CWS has been shown in the present study not to be predictive of success in the LCT, a verbal test low in assessment context.

Hypothesis 5.2 concerning the prediction of performance on the LCT from the HFT was tested next. Upon the addition of HFT scores in the MLR equation, $\Delta \underline{R}^2 = .01$, $\underline{F}(1,134) = 1.0$, $\underline{p} = .33$. Therefore the null hypotheses was not rejected. Contrary to expectations, extent of field independence as assessed by the HFT did not improve prediction of LCT performance beyond what was possible with the ability measures and physics achievement. The properties of the LCT items, irrelevant information in item stems, minimal redundancy of essential information, and diverse concepts among the response alternatives, were such that the formulation of tentative ideas of response requirements, the picking out of information from among the data of the item stem, and the reformulation of new hypotheses as suggested by the various response alternatives seemed likely to be required. These processes appeared to require the overcoming of both conceptual and perceptual embedding contexts and the ability to overcome embeddedness is the basis for making distinctions along the field independence dimension. The failure to observe the predicted overlap of extent of field independence and performance on the test low in assessment context may be due to a too-large gap between the conceptual nature of the embeddedness which had to be overcome in the LCT items as compared to the perceptual nature of the embeddedness to be overcome in the HFT. Witkin et al. (1962, chap. 5) has pictured the overcoming of embeddedness as being a cognitive trait having both conceptual and perceptual components. Extent of field independence is decidedly a part of the perceptual component. Possibly the overcoming of embeddedness involved in picking out data from an LCT item to test an hypothesis or in the formulation of a new response hypotheses after rejecting an initial one are so decidedly conceptual processes in their nature that the anticipated positive relationship between the LCT and the HFT did not exist.

The closeness of the relationship of the perceptual and conceptual realms is itself a matter of conjecture (Wohlwill, 1962). According to Wohlwill (1962), the Gestaltists tend to view perception as including insight and other processes which imply that the major part of human reasoning is perceptual while Bruner views the defining attributes of perception and conception as discontinuous. Wohlwill (1962) puts perception and conception at opposite poles of a continuum. As one proceeds from perception to conception the amount of redundant information needed for cognition decreases, the amount of irrelevant information which can be tolerated increases and the space-time separation over which stimulus information can be integrated increases. There are obvious parallels between the defining attributes of assessment context and distinguishing characteristics which Wohlwill (1962) employs along the perception - conception dimension.

If, for the moment, Wohlwill's (1962) formulation is accepted then a relatively simple explanation for the non-improvement of LCT prediction from the addition of the HFT follows, with the additional assumption that the making of responses to the LCT is decidedly a conceptual task. The conceptual nature of the LCT means that the lack of redundant information in the LCT, the presence of irrelevant information, and the diversity of the response alternatives poses no demand upon the perceptual facilities, while the making of responses to the HFT makes perceptual demands. Thus the capabilities assessed by the HFT and the LCT are separate.

A puzzling question is why was the HFT significantly predictive of scores on the HCT in which the items seemed to require little overcoming of embeddedness and not predictive of scores on the LCT in which the items seemed to require considerable overcoming of embeddedness? Possibly the ability to make rather fine distinctions among dominant elements in a test situation rather than the overcoming of embeddedness was the operative factor in the differing results for the HCT and the LCT. Wachtel (1972) has noted that tests used to assess field independence may be factorially complex. Hence the idea of the ability to make fine distinctions involving dominant elements cannot be immediately ruled out as not assessed by a test purporting to measure field independence. The ability to make fine distinctions involving dominant elements hypotheses could account for the significant relationship of the HFT with the HCT but not with the LCT since in the HCT the dominant elements of the items requirement was clearly spelled out in the item stem while for the LCT items it was not clearly defined in the stems. Furthermore, the single physics concepts among the response alternatives of the HCT may require that finer distinctions be made in the selection of responses than in the LCT where the differing physics concepts are similar to wide-spaced alternatives.

The same hypothesized component ability of field independence noted above seems to have further explanatory power. The superiority of highly field independent persons for learning lists of easily confused words (Long, 1962) and the closer relationship of field independence for tasks identifying the main elements of figures than for tasks identifying overall forms or backgrounds (Messick and Fritzky, 1963) may be accounted for on the basis of the ability to make fine distinctions involving dominant elements in the stimulus situation. The tendency of the highly field independent students to select preferentially mathematics and science courses (Wilkin, 1972) could also be explained if the assumption is warranted that mathematics and science courses involve the making of finer distinctions involving dominant elements in stimulus situations than is required in non-science and non-mathematics courses.

In summary, neither breadth of categorization or extent of field independence was found to improve the prediction of the low context verbal test, the LCT, beyond what was possible from physics achievement and the ability measures.

Tests with the DDT as the criterion. The hypotheses to be tested relate to MLR equations with physics achievement, breadth of categorization, and field independence as predictors. As for the HCT and the LCT tests, the hypotheses concerning physics achievement at the algorithmic thinking and knowledge levels have been combined with respect to physics achievement. Original hypotheses 3.3 and 3.6 follow as reformulated.

3.3 - .6 . There is no significant increase in the R^2 between the DDT and the ability variables when physics achievement is added to the set of predictors.

The hypotheses pertaining to breadth of categorization and extent of field independence follow:

4.3 There is no significant increase in the R^2 between the DDT and the ability variables plus physics achievement when breadth of categorization is added to the set of predictor variables.

5.3 There is no significant increase in the R^2 between the DDT and the ability plus physics achievement variables when the field independence variable is added to the set of predictor variables.

The results of the MLR tests are presented in Table 12.

Table 12

Tests of Hypotheses on Changes in the Squared Multiple Correlation of the DDT as the Criterion and Various Predictors

Step	Variables Added	R^2	ΔR^2	F	df	p	Decision on Variables
0	SCAT-V and SCAT-Q	.02	.02	1.5	1,136	.05	Covariates
1	P30	.37	.35	74.2	1,135	.01	Add
2	CWS	.38	.01	1.6	1,134	.22	Drop
3	HFT	.37	.01	.1	1,134	.75	Drop

Beat regression equation at $p < .05$:

$$X_{DDT} = -0.09X_{SCAT-V} - 0.07X_{SCAT-Q} + 0.66X_{P30}$$

The covariates, SCAT-V and SCAT-Q predicted a non-significant fraction of the variance of the DDT, $R^2 = .02$, $F(1,136) = 1.5$, $p > .05$. Probably the three year lag between the administration of the SCAT and the DDT was the reason for the weak association.

The addition of the P30 to the list of predictors resulted in an increase in the squared multiple correlation coefficient equal to 0.35, $F(1,135) = 74.2$, $p < .01$. Therefore null hypothesis 3.3 -.6 was rejected. As expected the physics achievement test was a significant predictor of the DDT undoubtedly because of the physics content which they have in common.

The addition of the CWS to the set of predictors yielded $\Delta R^2 = .01$, $F(1,134) = 1.6$, $p = .22$. Therefore, hypothesis 4.3 was not rejected. This means that breadth of categorization did not improve the prediction of performance on the diagrammatic test, the DDT. Such a result was anticipated because the DDT was essentially non-quantitative while significant relationships of the CWS have usually been found with quantitative tests (Pettigrew, 1958; Messick and Kogan, 1965). Furthermore, the examination of the diagrams of the DDT items for possible discrepancies between them and expectations based upon knowledge of physics is probably not a categorizing task.

The addition of the HFT to the set of predictors after the SCAT and P30, yielded an increase in the squared multiple correlation coefficient equal to 0.01, $F(1,134) = .1$, $p =$

.75. Therefore null hypothesis 5.3 was not rejected, which means that field independence was not found to aid in the prediction of performance on the DDT beyond what was possible from the ability tests and physics achievement. The result was contrary to expectations. Evidently the examination of the diagrams of the DDT for discrepancies between what is portrayed and what would be expected from a knowledge of physics, and the reassessment of expectations and the retesting of revised expectations against the given figures do not seem to be related to the process of picking out simple designs in complex figures as required in the HFT test of field independence. Possibly the diagrams of the DDT, not having the degree of complexity of the figures of the HFT, are not sufficiently challenging to provide the more field independent persons with an advantage over the less field independent students. There is another possibility. As discussed with reference to the LCT hypothesis tests, the conceptual component of the process of responding to the items of the DDT may be so much more important than the perceptual component that the relationship of the DDT and the HFT, a test with strong perceptual emphasis, is very weak.

Exploratory Analysis with the Assessment Context Tests as Criteria.

Cohen (1968) and Walberg (1971) have suggested that after testing well-developed MLR hypotheses exploratory variables including second order quadratic or interactions may be tested. Usually the exploratory variables are added to the MLR equations after the covariates and established variables. However, in the present study the quadratic terms and interaction terms have been generated and made available along with the original variables for a stepwise selection or rejection which has been used without restricting any variables from entering or being dropped from the MLR equation. The reason for following the procedure is that the covariates, SCAT-V and SCAT-Q were expected to play a very minor role in the prediction of the assessment context variables. Furthermore, the physics achievement test, the P30, was expected to play such a dominant role as a predictor that it seemed likely to appear in the regression equation formed by any stepwise process.

The method employed in testing the effects of second order quadratics and interactions among the predictors along with the original predictors is stepwise regression (Draper and Smith, 1966, chap. 6). In this method the predictor most highly correlated with the criterion is entered first into the regression equation and is retained if the improvement in prediction meets a certain pre-established probability level. Next the partial correlations between criterion and remaining predictors are calculated with the influence of the predictor already selected removed. The variable with the highest partial correlation is entered next and retained if it improves prediction at the pre-established level. Then, the variable(s) already in the equation are removed one at a time but are replaced if their removal reduces the predictability of the criterion at a pre-established level of probability. Then partial correlations of criterion and remaining predictors are calculated and the process continues.

The quadratic and interaction variables generated for inclusion in the stepwise process were:

$$X_{P30}^2 , \quad X_{CWS}^2 , \quad X_{HFT}^2 , \quad X_{P30} \cdot X_{CWS} ,$$

$$X_{CWS} \cdot X_{HFT} , \quad X_{HFT} \cdot X_{P30}$$

Also available for the selection by the stepwise process were the predictors:

$$X_{SCAT-V} , \quad X_{SCAT-Q} , \quad X_{P30} , \quad X_{CWS} , \quad X_{HFT} .$$

The probability level for variables entering or being dropped from the MLR equation was 0.05.

For the HCT as the criterion the following equation, in standard score form, resulted from the stepwise procedure:

$$X_{HCT} = 0.49 \; X_{P30}^2 + 0.24 \; X_{HFT} \cdot X_{P30}$$

The squared multiple correlation coefficient for the equation was equal to 0.47, $F(1, 136) = 59.9$, $p < .01$

For the LCT as the criterion, the following standard score equation resulted:

$$X_{LCT} = 0.61 \; X_{P30}$$

the squared multiple correlation coefficient was equal to 0.38, $F(1,137) = 83.9$, $p < .01$.

For the DDT as the criterion, the following standard score equation resulted:

$$X_{DDT} = 0.62\ X_{P30}$$

The squared multiple correlation coefficient was equal to 0.38, F(1,137)= 85.2, $p < .01$.

For the scores on the LCT and the DDT the P30 emerged from the stepwise procedure as the only significant predictor in each case.

For the HCT, two quadratic terms were each significant predictors. The two terms were the squares of scores on the P30, and the product of scores on the P30 and the HFT. The MLR equation shows that the P30 is the dominant element in that a person whose score on the P30 is zero would have a score of zero on the HCT, regardless of HFT performance. For zero scores on the HFT, increasingly high scores on the P30 produce even larger increases in HCT scores. For zero scores on the P30, zero scores occur on the HCT regardless of the size of HFT scores, according to the prediction equation.

The following partial derivatives of the HCT prediction equation with respect to the P30 variable and the HFT variable enable the relationships to be observed more clearly:

$$\frac{\partial X_{HCT}}{\partial X_{P30}} = 0.98\ X_{P30} + 0.24\ X_{HFT}$$

$$\frac{\partial X_{HCT}}{\partial X_{HFT}} = 0.24\ X_{P30}$$

Since the rate of change of HCT scores with respect to changes in HFT scores is directly proportional to P30 scores, while the rate of change in HCT scores with respect to P30 scores is directly proportional to P30 scores plus a linear factor proportional to HFT scores, a median split of the sample based on P30 scores would show greater divergence in HCT scores than would a median split of the sample based on HFT scores. If the meaning of the P30 is generalized to be considered a measure of ability in the discipline, and the HCT is taken to mean a specific performance within the discipline, then the following highly speculative hypothesis results. Median splits based upon ability are more effective for distinguishing between specific performances than are median splits based upon field independence. Such an hypothesis is in agreement with the finding that engineering grades were significantly predicted by an academic aptitude test and not by a field independence test (Brilhart and Brilhart, 1971). The hypothesis is not contradicted by the finding that science achievement of ninth year students was significantly predicted by a field independence test even though ability was controlled because only the extreme thirds of the sample based upon the field independence scores were considered in the prediction (Bowles and Boss, 1974).

Summary of the Results and Discussion

The strong association of the three assessment context tests with physics achievement was statistically significant as shown by zero order correlations among the variables, by canonical correlation and by MLR analysis. Such a result had been predicted based upon the common physics content of the assessment context tests and the achievement measure.

Breadth of categorization was found to be unrelated to any of the assessment context measures, although zero correlation coefficients significant beyond the .05 level were observed between the CWS and each of the ability measures. In the MLR analysis the CWS did not improve the prediction of any of the assessment context tests beyond what was possible from the ability measures and physics achievement. Calculation of the multiple correlation coefficient for each of the three assessment context tests, the HCT, the LCT and the DDT, with the SCAT-V, SCAT-Q and the P30 as predictors yielded R^2 values, in turn, of 0.41, 0.39, and 0.37. Recalculation of the R^2 for the HCT, LCT and DDT with only the P30 and CWS as predictors yielded values of 0.40, 0.38, and 0.36 respectively. Finally, R^2 values for the prediction of the HCT, LCT and DDT from the P30 alone yields, respectively, 0.40, 0.38, and 0.35. The comparison of the three sets of values reveals that the CWS is not quite as effective in conjunction with the P30 as are the SCAT-V and SCAT-Q with the P30 and furthermore that the addition of the CWS improves prediction of the assessment context tests minimally over what is possible from the P30 alone. Hence, whatever overlap exists between the ability measures and the CWS, the overlapping part is not effective in predicting achievement on the assessment context tests; if it were, approximately equal improvements in prediction would have resulted from the addition of the CWS or the SCAT-V and SCAT-Q to the P30 in the MLR equation.

The features which distinguished the test high in assessment context, the HCT, from the tests low in assessment context did not lead to the predicted relationships among the variances of the three tests. After the effects of physics achievement were removed, the partial correlation coefficients between the HCT and each of the LCT and DDT remained significantly positive, suggesting the presence of important similarities among the tests beyond that attributable to the physics content. For the HCT and the LCT the partial correlations could have been due to their similarity in the verbal, multiple choice format and in their items having been intermixed as part of a single test. But for the HCT and DDT the similarities did not exist; the HCT was verbal, the DDT diagrammatic; the HCT was administered at one time, the DDT a week later. An association of random variance of the HCT and the DDT may be the cause of the partial correlation. The K-R20 reliabilities of the HCT, LCT and DDT were, respectively, 0.49, 0.56, 0.61. Thus considerable random variance exists for each variable.

It is, however, also possible that the significant partial correlation between the HCT and each of the low context tests after the effects of physics achievement were removed may be due to the failure of assessment context to provide a basis on which to discriminate between tests. Conceivably the differences in the assessment context of the tests were not great enough. The differences could have been increased by increasing the redundancy of the essential information in the HCT items, by increasing the irrelevant information in the LCT and DDT items or by providing four or five different physics concepts among the response alternatives of the LCT. Whether such changes in the items are feasible without making the items unduly long or awkward remains an open question.

Field independence, as assessed by the HFT, was found to have significantly positive, $p < .05$, zero order correlation coefficients with the three assessment context tests and physics achievement. In the canonical analysis the first canonical correlation was interpreted as being primarily due to an association of physics achievement between the two variates; in addition the magnitude of the correlation between the HFT and the first predictor variate was 0.42 suggesting the importance of the HFT for predicting physics achievement. Multiple linear regression calculations for predicting separately each of the assessment context tests revealed that the HFT added significantly to the prediction of the HCT beyond what was possible from physics achievement and verbal and verbal and quantitative ability. The HFT did not however similarly improve the prediction of the

LCT and the DDT. Since the HCT was the one assessment context test constructed according to the criteria for a good classroom test, field independence appeared to play a role in the prediction of physics achievement as assessed by a multiple choice test constructed according to accepted procedures as outlined, for example, by Ebel (1972). The HFT items are characterized by the presentation of several items of information presented, often with redundancy in the item stem. The response alternatives usually have only one physics concept although the magnitude of the responses varies within each item.

An hypothesis which is suggested by the pattern of results for the prediction of the assessment context tests from the HFT is that where fine distinctions must be made among the dominant elements in a test situation the more field independent person has a decided advantage.

Exploratory analyses or a MLR equation employing second order quadratic and product terms among the predictors yielded no information on the prediction of the LCT and the DDT beyond what was obtained in the course of hypothesis testing. For the HCT, however, the following prediction equation emerged:

$$X_{HCT} = 0.49\ X_{P30}^{2} + 0.24\ X_{P30}\ X_{HFT}$$

The equation indicates the dominance of the P30 with progressively higher P30 values producing progressively larger increments in the HCT, and with the HFT operating in conjunction with the P30 to produce even larger increments in the HCT than would occur from increments in the P30 alone. The equation implies that field independence, although related to the test high in assessment context, would not be as discriminatory in predicting performance on the assessment context test as would the physics achievement test.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

The present chapter includes an overview of the purpose, hypotheses, method and results of the study. The implications for science education and for further research are presented as well as the limitations which must be considered in interpreting the results of the study.

Summary

Purpose. The purpose of the study was to examine the assessment context of test items in relation to physics achievement and two cognitive style variables, namely, extent of field independence and breadth of categorization. The assessment context of test items was postulated as an item property based on their similarity to certain characteristics of conventional test items in contrast to real-life situations. Assessment context was operationalized by the construction of physics tests varying in assessment context. The criteria for distinguishing between items high or low in assessment context were: amount of redundancy of the essential information in the item stem, the presence of irrelevant data in the item stem, and the number of different concepts among the response alternatives. The relationships between assessment context and ability and achievement variables were investigated for a sample of grade twelve physics students. Furthermore, assessment context was studied in relation to the cognitive styles of extent of field independence and breadth of categorization.

Major hypotheses. In order to accomplish the purpose of the study a question to be answered is the following one. Is the assessment context property of a test item of importance in the responses of students to the item? If assessment context is indeed a significant factor in the variance of achievement tests in, for example, physics, then a physics test with items which are high in assessment context should not have significant variance in common with a physics test with items low in assessment context once the effects attributable to knowledge of physics are removed. In particular, the following hypothesis was tested.

Within a given disciplinary field, two achievement tests which contrast in the assessment context of their items are not expected to be significantly related once the effects of knowledge of the discipline are removed.

Test items which are low in assessment context were postulated to contain more irrelevancies and potential hints at diverse explanations than are contained by items which are high in assessment context. The particular bits of information which are selected by the person responding to the items and the distinctions made between alternative avenues of explanation may be related to the individual's breadth of categorization. Breadth of categorization may be defined in terms of the number and type of properties of objects or stimuli the individual utilizes in sorting a collection of objects or in rating similarities among stimuli. Insofar as the irrelevancies and diverse hints in test items may be compared with the number and types of properties of objects or stimuli, the literature on breadth of categorization suggested the following hypothesis:

> Broad categorizers are likely to have the advantage over narrower categorizers in responding to items low in assessment context but not in responding to items high in assessment context.

Extent of field independence refers to the ease with which an individual is able to overcome an embedding context in a perceptual-conceptual cognitive task. Persons having a high degree of field independence are readily able to overcome embeddedness such as is required in locating a simple geometric pattern within a complex figure. Extent of field independence might be related to achievement on the assessment context tests because low context items have present irrelevant data and diverse concepts among the response alternatives while the high context items do not. The features of low context items appear to require the respondent to select specific subsets of the item data and to look for agreement between predictions from the data subsets and the various response alternatives. Insofar as the selecting of data subsets and examination of response alternatives are similar to overcoming embeddedness, the more field independent persons may be expected to be at an advantage in responding in test situations which are low in assessment context but not in situations which are high in assessment context. Thus the following hypothesis appeared to be warranted:

> Extent of field independence is significantly positively associated with performance on tests low in assessment context but is unrelated to achievement on tests high in assessment context.

Method. Two physics tests were prepared. One test of 40 items was multiple choice consisting of two subtests. One subtest had items high in assessment context and the other with low context items. The other test was constructed of 20 diagrammatic items having minimal verbal content and free response format. The test with diagrammatic items was considered to be a test low in assessment context.

In the construction and judging of the assessment context items the following three properties were considered: 1) the redundancy of the essential information in the item stem, 2) the extent of irrelevant information in the item stem, and 3) the diversity of concepts among the response alternatives for the test item. The physics items constructed for the study were rated high in assessment context if they contained two or more examples of redundant physics information, did not contain irrelevant physics information in the item stem, and had only one physics concept among the response alternatives.

The sample for the study comprised 144 grade 12 students of Physics 30 from three high schools in Edmonton, Alberta. Besides the diagrammatic test of items low in assessment context and the test consisting of subtests of items high and low in assessment context, five other tests were administered. The additional five tests assessed verbal and quantitative ability, extent of field independence, breadth of categorization and physics achievement.

Canonical correlation was used to assess overall relationships among the assessment context tests as criteria and the remaining five tests as predictors. Partial correlation was employed to calculate the strength of relationships among the assessment context tests and multiple linear regression analysis was used to delineate associations between each assessment context test as criterion and various combinations of predictors.

Results. The hypotheses on overall relationships between the assessment context variables as criteria and the ability variables, physics achievement, extent of field independence and breadth of categorization as predictors called for significant first and second canonical relationships. The strongest association between criteria and predictors was anticipated as due to the physics content of the context tests and the achievement measure and with the ability variables also contributing to the association. The anticipated relationship obtained, $p < .01$, with the three assessment context measures and physics achievement being the most important contributors to the first variate.

The second canonical correlation was anticipated to be significant because of linkages between the low assessment context variables and the cognitive style variables. However, the expected result was not found, $p = .4$. The failure to substantiate the hypothesis could not be attributed to the overwhelming strength of the first canonical relationship based mainly on the physics content since less than 50 percent of the variance of the criterion variables or predictor variables was accounted for in the first canonical correlation. The nonsignificant magnitude of the second canonical correlation suggests that field independence and breadth of categorization are not related to the assessment context variable independently of physics achievement.

The second set of hypotheses of the study concerned the relationship of the test high in assessment context and the ability, physics achievement, extent of field independence and breadth of categorization variables. It was hypothesized that the prediction of the high context variable is improved significantly upon the addition of physics achievement to the ability predictors but not significantly improved by the further addition of the cognitive style variables. The results of the multiple linear regression test showed that physics achievement was indeed significantly related to the high context variable. Such a result was not surprising because of the underlying physics content of the assessment context tests. However, the anticipated null result for the addition of the cognitive style variables was not achieved for extent of field independence. The field independence variable significantly improved, $p < .05$, upon the prediction of the high context variable beyond that which was possible from physics achievement and the ability variables. The unanticipated result was interpreted to mean that there is a perceptual-conceptual ability associated with the overcoming of an embedding context which contributes to success on tests of physics which are high in assessment context, that is, which are typical of conventional classroom tests. Further interpretation is possible in the light of the results of hypotheses concerning the prediction of the tests low in assessment context.

The hypotheses on the prediction of the low context tests called for improvement in prediction of each upon the addition of the physics achievement variable. Furthermore, extent of field independence was expected to improve the accuracy of predicting both the multiple choice subtest and the diagrammatic test. However, breadth of categorization was expected to improve upon the prediction of only the multiple choice subtest. The results of the tests of the hypotheses confirmed the effects of the physics achievement variable. The subsequent addition of the cognitive style variables improved the prediction neither of the multiple choice subtest nor of the diagrammatic test.

Extent of field independence which was anticipated as being positively related to success on the tests low in assessment context but unrelated to the high context test was found to behave in an opposite manner. It must be re-emphasized that the low context items were considered to resemble the nebulous real-life situations while the high context items were examples of conventional test items. The pattern of results for the field independence variable may be interpreted to mean that field independence plays a role in the way physics is taught and learned in most classrooms. The tests low in assessment context possessed features which seemed to place a premium upon the abilities to secure an overview of the item, to select subsets of the data, and to locate the response alternative which conforms to the conclusion obtainable from the data subset. The abilities just cited appear similar to those possessed by highly field independent persons. Yet extent of field independence was not predictive of success on the low context tests, which may be interpreted to mean that degree of field independence does not make its impact felt because of the type of test item. That is to say, the perceptual-conceptual requirement in responding to the test items does not have much impact on success on the test, at least insofar as the perceptual-conceptual requirements related to field independence are concerned. Possibly the minimal effect of the assessment context of the item is due to a decidedly conceptual requirement and a small perceptual component in the responding process.

If the failure to substantiate the hypotheses on the impact of the cognitive style variables in the low context situation is taken to mean that the assessment context of items does not affect the performance of individuals of varied cognitive styles, how is the improvement in prediction of the high context test by the addition of the field independence variable to be interpreted? The answer might be that in some way a high degree of field independence goes along with a good ability to learn physics.

Previous studies (Bowles and Boss, 1974; Barrett and Thornton, 1967) have indicated that highly field independent persons are high achievers in science but have not explained whether the superior achievement came about in the learning of the science or in the assessing of the achievement. The null results for the prediction of achievement on the low context tests of the present study imply that highly field independent persons do not have an advantage in responding to test items having features which seem to require complex responding processes. The inference may then be made that the higher achievement of the more field independent persons on the test high in assessment context and in science (Bowles and Boss, 1974) and engineering courses (Barrett and Thornton, 1967) is indeed due to factors involved in the learning of the science concepts. Such an inference also seems to be in agreement with the finding (Witkin, 1972) that highly field independent persons tend to select many optional science and mathematics courses in high school.

Conclusions. The concept of the assessment context of test items has been presented and investigated. Derived from an analysis of the characteristics present in real-life problem situations as compared with the problem presented in conventional classroom tests, assessment context seems to provide a basis for assessing the application-to-life objective within the limits of the classroom. Furthermore the idea of assessment context has been operationalized so that items low in assessment context, that is, similar to the real-life situation, may be generated. The need to be able to generate application items from operational definitions has been emphasized by Bormuth (1970). The availability of rules to operationalize the process of writing application items may be viewed as complementary to the methods for writing comprehension items (Anderson, 1972).

The investigation of assessment context as applied to physics test items has suggested that tests comprised of items low in assessment context may be constructed having difficulty levels and content validities comparing favorably with conventional classroom tests. Furthermore, the cognitive styles of breadth of categorization and extent of field independence do not appear related to achievement on the low context tests.

Implications for Science Education

1. The assessment context of test items has been based upon

an analysis of the characteristics of out-of-school problem situations as contrasted with the questions posed in conventional test items. Items classified as low in assessment context seem to have a logical validity for assessing the student's capability in the real-life situations whereas high context bear relatively little resemblance to the characteristics present in such situations. The concept of assessment context therefore provides science teachers with a new basis for judging test items. As operationalized in the present study, the assessment context of a science test item can be classified as low, high, or middle-valued in assessment context. Although items in physics have provided the basis for developing and studying assessment context, the basic criteria could readily be applied in other areas of science, or in non-science fields for that matter.

Application items, as defined by Bloom (1956), may be prepared readily by writing items low in assessment context. The "new slant" for testing application (Bloom, 1956, p. 125) can be provided by altering and enlarging the background of irrelevant information in which the necessary data is presented in the item stem, and by increasing the number of concepts among the various response alternatives. Ultimately however the full value of the assessment context concept cannot be realized until further studies have been done to measure the empirical validity of assessment context. The empirical validity will have been demonstrated when high achievement on low context items has been shown to accompany effective solving of real-life problems having the same underlying knowledge requirement. Concommitantly, ability in solving real-life problems must be shown to be less closely associated with ability in solving high context items than in solving low context items.

2. In order for a physics item to be low in assessment context irrelevant physics data must be present in the item stem, and two or more physics concepts must be present among the response alternatives. However, the presence of the preceding features means that the items are less than adequate as examples of "good" test items. Ebel (1972) has stated that test items should not contain "window dressing" and irrelevant information can be construed as such. Furthermore, the verb of the sentence containing the response alternatives or just preceding the response alternatives should be in the item stem, yet the multiple physics concepts among the response alternatives often demand different verbs for different alternatives. The nonstandard features of the low context items did not however appear to impair the criterion validity of the items since the low con-

text tests were at least as highly correlated with the physics criterion test as was the high context test. Furthermore, the reliabilities of the low context tests were at least as high as the high context test. The psychometric properties of the tests low in assessment context suggest that adequate criterion validity and internal consistency reliability may be achieved in tests consisting of items low in assessment context.

3. The results of the present study suggest that breadth of categorization is not related to physics achievement or to testing formats varying along the assessment context dimension. The literature on the nature of the categorization of objects according to classes established by the individual implies that breadth of categorization may be relevant to the capabilities of children in the concrete operations stage of development; however, studies of the relationship do not appear to exist.

4. The connection of field independence with physics achievement and assessment context is less clear than that of breadth of categorization. Highly field independent students exhibit superior performance in science and mathematics although the reports in the literature do not always agree on whether or not the positive association is statistically significant. Doubt remains as to whether the positive association derives from

factors present in the achievement measurement process, the teaching-learning activities in science and mathematics, or a pre-existing aptitude for science and mathematics of highly field independent persons. The results of the present study indicate that the positive association may be related to the characteristics of the measuring instrument, at least for physics, since field independence emerged as a significant predictor of the High Context Test but not of the Low Context Test and the Discrepancy Detection Test. The literature suggests however, that the positive association may be related to the method of instruction. Given the possibility that measured physics achievement may be related to extent of field independence as well as to knowledge of physics then, in fairness to all students, teachers should remember Nedelsky's (1965, p. 104) advice with respect to testing; namely, a range of item types is likely to be fairer than reliance upon a single type. The instructional counterpart of Nedelsky's advice on testing would be: The fairest approach is to employ a variety of instructional methods rather than to rely upon a single one. The preceding recommendation recognizes that Messick's (1970) claim that cognitive styles play unknown roles in schools is yet valid, at least for field independence. Therefore, the guidance of students entering senior high school into particular course patterns based upon measured field independence would be premature.

Implications for Further Research

1. Further investigation of the contention that the low context items bear a closer resemblance to real-life applications of knowledge than do high context items is required to assess the empirical validity of the claim. The continued stress in the science education literature on the importance of the application-to-life objective emphasizes the need to be able to assess achievement of the objective while students are in school.
2. The nature of the relationship of breadth of categorization and quantitative ability requires elucidation. The original 20-item scale for assessing breadth of categorization was found to be bifactorial, with one factor positively associated with quantitative ability (Pettigrew (1958). However, Messick and Kogan (1965) found only the other factor to be related to quantitative ability and then only for ability tests with wide-spaced alternatives. The 10-item abbreviated form of the scale, the CWS, was found in the present investigation to be significantly associated with quantitative ability. Being able to generalize about the relationship could have benefits for understanding the way students approach problems involving quantitative calculations.

3. Further research is required to clarify the possible relationship of field independence and the science achievement measurement process. The clarification is required lest the achievement measurement factors be confounded with instructional method factors in the overall relationship of science performance and field independence. Reports in the literature suggest that field independence and achievement in mathematics and science are associated. The results of the present investigation showed that extent of field independence improved the prediction of achievement or the High Context Test but not on the Low Context Test or on the Discrepancy Detection Test. The findings imply that the association may be due to factors present in the measurement process, at least in the measurement of achievement in physics.
4. The relationships among the various tests of field independence need to be studied further. Taken together the results of many reports from the literature on field independence suggested that with adult subjects sex differences in field independence are found with the Rod and Frame Test but not with the Hidden Figures Test. Furthermore, field independence is not often found related to general intellectual ability when the Rod and Frame Test measure is involved but a significant positive association is frequently found when the

Hidden Figures Test or the Embedded Figures Test is used to assess field independence. It is possible that the pattern outlined is due to sampling factors. Whatever the case a deeper understanding is required of field independence and the measures used to assess it.

5. Diagrammatic items for assessing physics achievement need to be explored further with respect to validity and efficiency. Diagrammatic items comprised the Discrepancy Detection Test of the present study and the Test was as highly correlated with physics achievement as were the verbal assessment context tests. In addition less time per item was required in responding to the diagrammatic items as compared with the verbal items.
6. Studies similar to the present one are desirable in the other sciences and in mathematics. High and low assessment items could be pushed to greater extremes in assessment context and the subject matter content could cover a more limited range of concepts. The more limited range would enable the tests to be administered earlier in the school year and the greater extremes of assessment context tests should provide clearer evidence of the role of assessment context.

Limitations of the Study

The method of the study was carried out within certain limitations which must be considered in interpreting the findings. The assessment context tests were restricted to lengths of 20 items each because of limits on the amount of class time which could be spent on testing. The shortness of the tests tended to make their reliability less than is desirable. Another limitation within the study is in the amount of difference in assessment context of the items of the LCT and the HCT. The differences could have been greater, but were not, in order that the students would not become aware that the LCT items were strange or unusual. One of the distinguishing points of assessment context was that items of the LCT had to contain irrelevant physics information. If only one datum of irrelevant information was present the criterion was considered met, but the criterion could have been changed so that, for example, five bits of irrelevant physics data were required. Similarly, the requirement that low context items had to contain more than one physics concept among the response alternatives could have been changed to require more than three physics concepts among the response alternatives. However, the forcing of more extreme requirements into the items of the LCT would probably have made the items longer than the HCT items and have made the low context items noticeably different in the eyes of the students thereby inducing unpredictable response styles.

The Discrepancy Detection Test, the DDT, is a test of unproven quality. While other diagrammatic tests have been used in physics they are multiple choice, not free response as is the DDT, and they do not ask the student, as the DDT does, to look for contradictions between what is portrayed in the diagrams and what is to be expected based upon a knowledge of physics.

Inherent in the use of canonical correlation and multiple linear regression equations is the danger that the optimization calculations may capitalize on chance factors (Meredith, 1964; Cohen, 1968; Draper and Smith, 1966; Cooley and Lohnes, 1971). The danger is especially great where the sampling of students and achievement test items are non-random as in the present study. Procedures followed to minimize the effect of chance factors include the interpretation of the canonical correlations in terms of the structural coefficients rather than in terms of the factor weights for each variable. In addition, the redundancy index of Stewart and Love (1968) for describing the efficiency of prediction of the canonical variates is an index which considers all of the predictor or criterion variables together and which therefore avoids placing emphasis on idiosyncrasies associated with any one variable. The multiple linear regression equations have been interpreted in terms of the multiple correlation coefficients rather than with respect to the weights applied to particular variables. In spite of the precautions outlined above, it must be emphasized that the student sample was not a random one, and that confidence in multiple linear regression and other optimization techniques comes by replication rather than by reducing the acceptable level of probability of a type one error.

Because of considerations involving the schools, the tests were not administered under a counter-balanced plan. Hence, certain sequencing effects may have been present.

Another limitation of the study is in the assessment of extent of field independence by means of a single instrument. Because of apparent inconsistencies in what is measured by several of the instruments for assessing field independence, Vernon (1972) has recommended the administration of a small battery of tests including the Rod and Frame Test, the Embedded Figures Test, and the Draw A Person Test. Presumably the first principal component of the battery would be accepted as the estimate of the field independence construct. The disruption of student school schedules which results from the administration of individual tests made unfeasible the following of Vernon's suggestion for the present investigation and forced reliance upon the group-administered HFT as the sole measure of extent of field independence.

FOOTNOTES

1. The Hidden Figures Test has not been presented in the report at hand at the request of Educational Testing Service.
2. The Category Width Scale is presented in appendix F.
3. The Physics 30 Test has not been reproduced at the request of the Alberta Department of Education.

REFERENCES

Anderson, C. C. & Cropley, A. J. Some correlates of originality. *Australian Journal of Psychology*, 1966, *18*, 218-227.

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, *42*, 145-170.

Anderson, T. W. *An introduction to multivariate statistical analysis*. New York: Wiley, 1958.

Attneave, F. *Application of information processing to psychology*. New York: Holt, 1959.

Avital, S. M. & Shettleworth, S. J. *Objectives for mathematics learning*. Toronto: The Ontario Institute for Studies in Education, Bulletin No. 3, 1968.

Barrett, C. N. & Thornton, C. Cognitive style differences between engineers and college students. *Perceptual and Motor Skills*, 1967, *25*, 789-793.

Bartlett, M. S. Multivariate analysis. *Supplement to the Journal of the Royal Statistical Society*, 1947, *9*, 176-197.

Bieri, J. Category width as a measure of discrimination. *Journal of Personality*, 1969, *37*, 513-521.

Bieri, J., Bradburn, W. M. & Galinsky, M. P. Sex differences in perceptual behavior. *Journal of Personality*, 1958, *26*, 1-12.

Bloom, B. S. *Taxonomy of educational objectives: Cognitive domain*. New York: David McKay, 1956.

Bloomberg, M. Field-independence-dependence and susceptibility to distraction. *Perceptual and Motor Skills*, 1965, *20*, 805-813.

Board, C. & Whitney, D. The effects of selected poor item-writing practices on test difficulty, reliability and validity. *Journal of Educational Measurement*, 1972, *9*, 225-233.

Bock, R. D. & Haggard, E. A. The use of multivariate analysis of variance in behavioral research. In D. R. Whitla (Ed.), *Handbook of measurement and assessment in the behavioral sciences*. New York: Addison-Wesley, 1967.

Boersma, F. J. Test-retest reliability of the CF-1 hidden figures test. Educational and Psychological Measurement, 1968, 28, 555-559.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Bowles, A. & Boss, M. W. Extent of psychological differentiation as related to achievement in science and attitude toward science. Paper presented at the American Educational Research Association, Annual Meeting, Chicago, 1974.

Brilhart, B. L. & Brilhart, J. K. Field independence and academic achievement of engineering students. Perceptual and Motor Skills, 1971, 32, 443-446.

Bruner, J. S., Goodnow, J. & Austin, G. A. A study of thinking. New York: Wiley, 1956.

Bruner, J. S. & Tajfel, H. Cognitive risk and environmental change. Journal of Abnormal and Social Psychology, 1961, 62, 231-241.

Buros, O. K. (Ed.). The sixth mental measurements yearbook. Highland Park, N. J.: Gryphon, 1965.

Cattell, R. B. & Butcher, H. J. The prediction of achievement and creativity. Indianapolis: Bobbs-Merrill, 1968.

Cohen, J. Some statistical issues in psychological research. In B. B. Wolman (Ed.), Handbook of clinical psychology. New York: McGraw-Hill, 1965.

Cohen, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Cooley, W. W. & Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Davis, J. K. Strategy development and hypothesis testing as a function of an individual's cognitive style. Final Report to Office of Educational Research, Washington, D. C., 1972. Grant No. OEG-5-71-0035 (509). (ERIC Document Reproduction Service No. ED 013 371).

Davis, J. K. Cognitive style and hypothesis testing. Paper presented at the meeting of the American Educational Research Association, New Orleans, 1973. (ERIC document Reproduction Service No. ED 072 388).

Davis, J. K. & Klausmeier, H. J. Cognitive style and concept identification as a function of complexity of training procedures. Journal of Educational Psychology, 1970, 61, 423-430.

Denmark, F. L., Havlena, R. A. & Murgatroyd, D. Reevaluation of some measures of cognitive styles. Perceptual and Motor Skills, 1971, 33, 133-134.

DeRussy, E. A. & Futch, E. Field-independence-dependence as related to college curricula. Perceptual and Motor Skills, 1971, 33, 1235-1237.

Draper, N. R. & Smith, H. Applied regression analysis. New York: Wiley, 1966.

Dressel, P. L. & Schmid, J. Some modifications of the multiple-choice item. Educational and Psychological Measurement, 1953, 13, 574-595.

Dreyer, A. S., Nebelkopk, E., & Dreyer, C. A. Note concerning stability of cognitive style measures in young children. Perceptual and Motor Skills, 1969, 28, 933-934.

Dubois, T. F. & Cohen, W. Relationship between measures of psychological differentiation and intellectual ability. Perceptual and Motor Skills, 1970, 31, 411-416.

Dunn, T. F. & Goldstein, L. G. Test difficulty, validity and reliability as functions of selected multiple-choice item construction principles. Educational and Psychological Measurement, 1959, 19, 171-179.

Eagly, A. H. Responses to attitude-discrepant information as a function of tolerance of inconsistency and category width. Journal of Personality, 1969, 37, 601-617.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Elliot, R. Interrelationships among measures of field dependence, ability and personality traits. Journal of Abnormal and Social Psychology, 1961, 63, 27-36.

Feather, N. T. Level of aspiration and performance variability. *Journal of Personality and Social Psychology*, 1967a, *6*, 37-46.

Feather, N. T. Some personality correlates of external control. *Australian Journal of Psychology*, 1967b, *19*, 253-260.

Field, T. W. & Cropley, A. J. Cognitive style and science achievement. *Journal of Research in Science Teaching*, 1970, *7*, 2-10.

Fillenbaum, S. Some stylistic aspects of categorizing behavior. *Journal of Personality*, 1959, *27*, 187-195.

Fleishman, J. J. & Dusek, R. Reliability and learning factors associated with cognitive tests. *Psychological Reports*, 1971, *29*, 523-530.

French, J. W., Ekstrom. R. B., & Price, L. A. *Kit of Reference Tests for Cognitive Factors*. Princeton, N. J.: Educational Testing Service, 1963.

Gagne, R. M. *The conditions of learning* (2nd ed.). New York: Holt, Rinehart & Winston, 1970.

Gardner, R. W. Cognitive styles in categorizing behavior. *Journal of Personality*, 1953, *22*, 214-233.

Gardner, R. W., Holzman, P. S., Klein, G. S., Linton, H. & Spence, D. Cognitive control. *Psychological* Issues, 1959, *1*, Monograph 4.

Gardner, R. W., Jackson, D. N. & Messick, S. Personality organization in cognitive controls and intellectual abilities. *Psychological Issues*, 1960, *2*, Monograph 8.

Gardner, R. W. & Schoen, R. A. Differentiation and abstraction in concept formation. *Psychological Monographs*, 1962, *76*, Whole No. 560.

George, J. D. Relation of selected aptitudes to the verbal and diagrammatic content of physics tests. M.Ed. thesis, The University of Alberta, Edmonton, 1971.

Green, R. The SCAT. In O.K. Buros, *The sixth mental measurements yearbook*. Highland Park, N. J.: Gryphon, 1965, 452-453.

Grieve, T. D. & Davis, J. K. The relation of cognitive style and method of instruction to performance in ninth grade geography. Journal of Educational Research, 1971, 65, 137-141.

Guttman, L. A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), Mathematical thinking in the behavioral sciences. Glencoe, N. Y.: The Free Press, 1954.

Hedges, Wm. D. Testing and evaluation for the sciences. Belmont, Calif.: Wadsworth, 1966.

Highley, F. S. Verbal ability, quantitative ability and the rod-and-frame test. Perceptual and Motor Skills, 1970, 30, 957-958.

Hotelling, H. Relations between two sets of variates. Biometrika, 1936, 28, 321-377.

Hudson, L. Contrary imaginations. London: Methuen, 1966.

Hunt, E. B. Concept learning: An information processing problem. New York: John Wiley and Sons, 1962.

Hurd, P. D. Integrated science. The Science Teacher, 1973, 40, 18-19.

Jackson, D. N., Messick, S. & Myers, C. T. Evaluation of group and individual forms of embedded-figures measures of field-independence. Educational and Psychological Measurement, 1964, 24, 177-192.

Jackson, R. H. & Pedersen, D. N. Some correlates of mode of conflict resolution in impression formation. Perceptual and Motor Skills, 1965, 21, 635-644.

Karp, S. A. Field-independence and overcoming embeddedness. Journal of Consulting Psychology, 1963, 27, 294-302.

Klein, G. S. Personality. Annual Review of Psychology, 1967, 18, 467-560.

Kogan, N. & Wallach, M. A. Risk taking. New York: Holt, Rinehart and Winston, 1964.

Kropp, R. P., Stoker, H. W. & Bashaw, W. L. The validation of the taxonomy of educational objectives. The Journal of Experimental Education, 1966, 34(3), 69-76.

Lester, G. Comparison of five methods of presenting the rod-and-frame test. Perceptual and Motor Skills, 1969, 29, 1947-151.

Lewis, D. G. Objectives in the teaching of science. Educational Research, 1965, 7, 186-199.

Lezotte, W. L. The relationship between cognitive styles, scholastic ability, and the learning of structured and unstructured materials. Unpublished Doctoral Dissertation, Michigan State University, 1969. (University Microfilms No. 70-9587).

Loevinger, J., Gleser, G. C. & DuBois, P. H. Maximizing the discriminating power of a multiple-score test. Psychometrika, 1953, 18, 253-262.

Long, R. I. Field articulation as a factor in verbal learning and recall. Perceptual and Motor Skills, 1962, 15, 151-158.

McMorris, R. F., Brown, J. A., Snyder, G. W. & Pruzek, R. M. Effects of violating item construction principles. Journal of Educational Measurement, 1972, 9, 287-295.

McNemar, Q. Psychological statistics (4th edition). New York: Wiley, 1969.

McQuitty, L. Improved hierarchical syndrome analysis of discrete and continuous data. Educational and Psychological Measurement, 1966, 26, 577-582.

Meredith, Wm. Canonical correlations with fallible data. Psychometrika, 1964, 29, 55-65.

Messick, S. The criterion problem in the evaluation of instruction. In M. C. Wittrock and D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart and Winston, 1970.

Messick, S. & Damarin, F. Cognitive style and memory for faces. Journal of Abnormal and Social Psychology, 1964, 69, 313-318.

Messick, S. & Fritzky, F. J. Dimensions of analytic attitude in cognition and personality. Journal of Personality, 1963, 31, 346-370.

Messick, S. & Kogan, N. Category width and quantitative aptitude. Perceptual and Motor Skills, 1965, 20, 493-497.

Miller, J. K. The development and application of bi-multivariate correlation: A measure of statistical association between multivariate measurement sets. Doctoral Dissertation, Faculty of Educational Studies, State University of New York at Buffalo, 1969. (University Microfilms No. 69-20595).

Nedelsky, L. Science teaching and testing. New York: Harcourt, Brace and World, 1965.

Penk, W. Two measures of category width: Age, sex, examiner differences and intercorrelations. Psychological Reports, 1969, 25, 859-870.

Pettigrew, T. F. The measurement and correlates of category width as a cognitive variable. Journal of Personality, 1958, 26, 532-544.

Podell, J. E. & Phillips, L. A developmental analysis of cognition as observed in dimensions of Rorschack and objective test performance. Journal of Personality, 1959, 27, 439-463.

Podrasky, E. F. Nonverbal assessment of learning. The Science Teacher, 1971, 38(6), 39-41.

Poole, R. L. Characteristics of the taxonomy of educational objectives: Cognitive domain. Psychology in the Schools, 1971, 8, 379-385.

Pysh, F. The relationship of field-independence to performance on Piagetian-type tasks incorporating the Euclidean coordinate system. The Western Psychologist, 1970, 1, 137-143.

Rosen, M. Post decision affinity for incompatible information. Journal of Abnormal and Social Psychology, 1961, 63, 188-190.

Rowe, M. B. Influence of context-learning on solution of task oriented science problems. Journal of Research in Science Teaching, 1965, 3, 12-18.

SCAT Technical Report, Cooperative School and College Ability Tests. Princeton: Educational Testing Service, 1958.

SCAT-STEP Technical Supplement. Cooperative School and College Ability Tests. Princeton: Educational Testing Service, 1962.

Schwartz, D. W. & Karp, S. A. Field-independence in a geriatric population. Perceptual and Motor Skills, 1967, 24, 495-504.

Sherman, J. Problem of sex differences in space perception and aspects of intellectual functioning. Psychological Review, 1967, 74, 290-299.

Sloane, H. N., Gorlow, L. & Jackson, D. N. Cognitive styles in equivalence range. Perceptual and Motor Skills, 1963, 16, 389-404.

Smith, E. R. & Tyler, R. M. Appraising and recording student progress. New York: Harper and Brothers, 1942.

Smith, I. L. Validity of taxonomic tests. Educational and Psychological Measurement, 1971, 31, 475-476.

Smith, R. B. An empirical investigation of complexity and process in multiple-choice items. Journal of Educational Measurement, 1970, 7, 33-42.

Sommerfield, R. E. & Tracy, N. H. A study of selected predictors of success in second year algebra in high school. The High School Journal, 1963, 46, 234-240.

Spotts, J. J. & Mackler, B. Relationship of field-dependent and field-independent cognitive styles to creative test performance. Perceptual and Motor Skills, 1967, 24, 239-268.

Stewart, D. & Love, Wm. A general canonical correlation index. Psychological Bulletin, 1968, 70, 160-163.

Stoker, H. W., & Kropp, R. P. An empirical validity study of the assumptions underlying the structure of cognitive processes using the Guttman-Lingoes smallest space analysis. Educational and Psychological Measurement, 1971, 31, 469-473.

Stollberg, R. & Hill, F. Fundamentals of physics (Canadian edition). New York: Houghton and Mifflin, 1968.

Stuart, I. R. Perceptual style and reading ability: Implications for an instructional approach. Perceptual and Motor Skills, 1967, 24, 135-138.

Stuart, I. R. and Bronzaft, A. L. Perceptual style, test anxiety and test structure. Perceptual and Motor Skills, 1970, 30, 823-825.

Tatsuoka, M. M. & Tiedeman, D. V. Statistics as an aspect of scientific method in research on teaching. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Tempero, H. E. & Ivanoff, J. M. The cooperative school and college ability test as a predictor of achievement in selected high school subjects. Educational and Psychological Measurement, 1960, 20, 835-838.

Thornton, C. L. & Barrett, G. V. Methodological note on N-achievement and field-independence comparisons. Journal of Consulting Psychology, 1967, 31, 631-632.

Thurstone, L. L. A factorial study of perception. Chicago: University of Chicago Press, 1944.

Vaught, G. M. The relationship of role identification and ego strength to sex differences in the rod-and-frame test. Journal of Personality, 1965, 33, 271-283.

Vernon, P. E. The distinctiveness of field independence. Journal of Personality, 1972, 42, 366-391

Wachtel, P. L. Style and capacity in analytic functioning. Journal of Personality, 1968, 36, 202-212.

Wachtel, P. L. Cognitive style, attention and learning. Perceptual and Motor Skills, 1971, 32, 315-318.

Wachtel, P. L. Field-independence and psychological differentiation: Reexamination. Perceptual and Motor Skills, 1972, 35, 179-189.

Walberg, H. J. Generalized regression models in educational research. American Educational Research Journal, 1971, 8, 71-91.

Wallach, M. A. & Caron, A. J. Attribute criteriality and sex-linked conservatism as determinants of psychological similarity. Journal of Abnormal and Social Psychology, 1959, 59, 43-50.

Wertheimer, M. Productive thinking. New York: Harper and Brothers, 1945.

Witkin, H. A. Individual differences in ease of perception of embedded figures. Journal of Personality, 1950, 19, 1-15.

Witkin, H. A. The role of cognitive style in academic performance and in teacher-student relations. Paper presented at the Symposium sponsored by the Graduate Record Examination Board, Montreal, 1972. (ERIC Document Service No. ED 083 248).

Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R. & Karp, S. Q. Psychological differentiation. New York: John Wiley, 1962.

Witkin, H. A., Goodenough, D. R. & Karp, S. A. Stability of cognitive style from childhood to young adulthood. Journal of Personality and Social Psychology, 1967, 7, 291-300.

Witkin, H. A., Lewis, L. B., Hertzman, M., Machover, K., Meissner, P. B. & Wapner, S. Personality through perception. New York: Harper and Brothers, 1954.

Wohlwill, J. F. From perception to inference: A dimension of cognitive development. Monographs of the Society for Research in Child Development, 1962, 27, 87-112.

Wood, J. A., Allers, R. A., Hornsby, J. A., Redcliffe, L., Westbury, R. & White, S. A. Statistical supplement to summary description of grade nine literature objectives, test items and blueprint. Edmonton: Department of Education, 1968.

APPENDIX A

Instructions for Rating Assessment
Context

INSTRUCTIONS TO READERS

1. Please read the enclosed "Instructions for the Evaluation of Item Context".

2. Rate each of the 40 items on the three context variables described therein, and record your rating of each one beside the appropriate item number on the enclosed "Rating Sheet".

3. Beside each of the test items write "Valid" or "Invalid" based upon whether or not you consider the item samples content of the Physics 30 Course studied by pupils. Any comments on possible ambiguities, or other notes which you might make would be appreciated.

INSTRUCTIONS FOR THE EVALUATION OF ITEM CONTEXT

Introduction

For purposes of a study on the relationship of test item properties and pupil performance, the term "item context" has been defined.

In order to provide data for assessing the validity of "item context" you are asked to read the explanation of it which follows, and to rate a series of items as being high or low in "item context", sometimes referred to in terms of "context value".

Explanation

The context value of a multiple choice item is related to the following properties: (1) the amount of redundancy in the information essential for correctly solving the item, (2) the presence of information in the item stem which is irrelevant to solving the item, and(3), the diversity of the concepts to be found among the various alternative responces provided in the item.

Redundancy of the essential information in a test item is evident when a diagram is used to repeat information given verbally, or when excess information is provided by repetition of concepts by symbols, or by naming the originator of a law of physics as well as stating the law, or by providing measurement unit names as well as the names of the quantities being measured. Consider the following contrasting examples:

A. Newton's second law states that a mass of (M) kg will accelerate at (a) m/sec^2 when acted upon by an accelerating force of (F) newtons. If the force is increased to (2F) newtons, and the mass is decreased to (½M) kg, the acceleration will become

1. (4a) m/sec^2
2. (2a) "
3. (a) "
4. (½a) "
5. (¼a) " .

B. When a mass is acted upon by an unbalanced force, it will accelerated. The magnitude of the acceleration is directly proportional to

1. the mass and to the force
2. the force and to the square of the mass
3. the mass and inversely proportional to the force
4. the product of mass and force
5. the force and inversely proportional to the mass.

Essentially the same information is conveyed in A and in B but with less redundancy in B. Item A would be rated higher in context value than item B.

The second element which affects context value is the presence of information irrelevant to that required for solving the item. When irrelevant information is present in an item, the item is rated as low in context value. The following items illustrate the presence or absence of irrelevant information:

C. A wooden cube has a mass of 1.0 slug. The cube is 1.2 ft on each edge; its specific gravity is 0.29. An unbalanced force of 10 lbs accelerates the cube through a distance of 20 ft. The acceleration of the cube is

1. 24 ft/sec^2
2. 10 "
3. 0.31 "
4. 200 "

D. A mass of 1.0 slug is acted upon by an unbalanced force of 10 lbs. The acceleration of the object is

1. 100 ft/sec^2
2. 10 "
3. 0.10 "
4. None of the above answers.

Item C would be rated as not containing only information essential to solving the item, whereas item D contains only information, possibly redundant information, which is needed for answering it.

The third property on which item context depends is the diversity of concepts presented among the alternative responses to an item. If all of the responses to an item are related to the same concepts, for example, all kinetic energy amounts, the item is of higher context value than if each of the alternative responses demands the consideration of separate concepts. The following items are provided for comparison:

E. Two objects, one having a mass (m), the other a mass (2m), are simultaneously released from the top edge of a tall building. By the time that each mass has fallen 50 ft, the larger mass will have

1. a kinetic energy 4 times that of the smaller
2. " " " the same as " " " "
3. " " " ½ " " " "
4. " " " " " " ".

F. Two objects, one having a mass (M) and density (D_1), the other a mass (2M) and density (D_2), are simultaneously released from the top edge of a 100 ft high building. By the time the two masses have fallen 50 ft, the larger mass will have

1. the same kinetic energy as the smaller mass
2. " " momentum " " " "
3. " " net force acting " " " "
4. " " velocity " " " ".

Because the the wide diversity of concepts in the alternative responses to item F as compared to item E , and for other reasons too, item F would be rated lower in context than item E.

Item context, then, is related to the amount of redundancy of essential information, the amount of information in the item stem which is irrelevant to answering the element, the extent of the diversity of concepts required in choosing among the alternative responses. An item having redundancy of the essential information, no irrelevant information, and but one concept among the response alternatives would certainly be rated as high in assessment context. But what of items having only one or two of these properties?

It would be possible to write test items with all possible combinations of the three properties described above. The number of possible combinations of these three properties in a test item would be eight , based upon redundancy of the essential information (Yes or No), only relevant information (Yes or No), and single concept response alternatives (Yes or No). Hence the number of possible combinations is 8=2x2x2. In the preceding test items used for illustration, not all of the eight combinations were to be found.

Shown below is a table with various combinations of test item variables and the assessment context decision rule for each combination. (Y) means Yes; (N) means No.

Some combinations of item variable ratings are considered to be of intermediate context value and hence are not rated in the "Context Decision" column.

If a test item contains redundancy of the information essential for solving the item it is rated (Y) for that category regardless of the extent of the redundancy.

In a similar way, where an item contains information irrelevant to reaching a correct solution the item will be rated as (N) in the category "Only relevant information present in item stem" regardless of whether one, two, or more pieces of irrelevant information are present. Likewise, if more than one concept is shown to be required in the various response alternatives, the item is rated (N) in the category "Single concept response" no matter how many different concepts are presented among the alternatives.

Item	Redundancy of Essential Information	Only Relevant Data Present in item stem	Single Concept Response Alternatives	Context Decision
A, D, E	Y	Y	Y	High
	Y	Y	N	
	N	Y	Y	
B	N	Y	N	
C	Y	N	Y	
	N	N	Y	
F	Y	N	N	Low
	N	N	N	Low

FIGURE 1. Examples of item classifications and context decisions.

Summary

An item is rated (Y) for redundancy if it contains an excess of the essential information in repeating verbal information by means of a diagram, naming a physics concept, for example, mass, force, energy etc. and also providing a symbol, for example, M, F, E, and/or giving units of measurement, for example, kg, newtons, joules. Although items can have varying amounts of redundancy, even one instance of the above forms of redundancy is sufficient for a rating of (Y).

The presence of verbal information about physics which is irrelevant to solving the item means that under the column headed "Only Essential Information Present" would be recorded (N). Emphasis is made of the fact that in rating an item in this category, only the stem of the item is to be considered.

The presence of more than one physics concept among the response alternatives would mean that an (N) would be recorded under the appropriate category rather than a (Y).

Two further examples of items and their ratings are provided below:

G. A cylindrical object has a radius of 1.0 unit, a height of 2.0 units and a weight of 50 lbs. The volume of this object is

1. 50 cubic units
2. $\pi \times (1)^2 \times 2$ cubic units
3. 2.0 cubic units
4. 4.0 cubic units .

H. A cylindrical object has a radius of 1.0 ft, a height of 2.0 ft and a weight (Wt) of 50 lbs. The density (D) of this object is

1. greater than that of water
2. equal to that of water
3. less than that of water
4. not determinable from the information given.

Item G would be rated

N N Y

since there is no redundancy in the information provided which is essential to finding the volume of the object. This accounts for the first N. The weight information is an irrelevancy, not a redundancy of essential information; hence the second N. Only one concept, that of volume, is to be found necessary among the response alternatives, hence the Y-rating. This particular pattern, N N Y , is shown in Table 1 to be descriptive of a Low context item .

Item H would be rated

Y Y Y High

since there is redundancy of the information needed to find density in that "radius","height", "weight" are each designated with units of length, length, and weight respectively. Also the symbols (Wt) for weight and (D) for density are given. No irrelevant information is present, and only one concept of physics, that of density comparison, is required among the response alternatives.

ITEM RATING SHEET

Item Number	Redundancy of Essential Information	Only Relevant Data Present in Item Stem	Single Concept Response Alternatives	Context Decision
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
33				
34				
35				
36				
37				
38				
39				
40				

APPENDIX B

The Physics Test

PHYSICS TEST

This is a test of understanding of physics.

Each test item provides 5 possible choices of answer. Select the one you think is best and darken the space corresponding to it on the answer sheet.

Some items in this test contain information which is not essential for finding the best answer. Care in reading such items will be necessary in order to decide exactly what is required.

Assume that there is no friction or air resistance in any of the situations described unless that test item specifically states that such forces exist.

Should it be required, the acceleration of gravity at the earth's surface is 9.8 m/sec^2, or 32 ft/sec^2.

Two sample items are given below, one with irrelevant information, the other without unnecessary information.

Sample items:

1. A wooden cubical block, 2 ft. on each edge, has a weight of 160 lbs. The volume of this block is

   A. 160 cu ft
   B. 2 cu ft
   C. 4 cu ft
   D. 8 cu ft
   E. not determinable from the information given

2. The pound is a unit for measuring

   A. weight
   B. energy
   C. momentum
   D. impulse

E. power.

Sample answer sheet:

1. A B C [✓ D] E

2. [✓ A] B C D E

1. A box of mass 2.0 kg has a speed of 4.0 m/sec. The kinetic energy of the box is

   A. 8.0 J
   B. 16.0 J
   C. 78.4 J
   D. 2.0 J
   E. not determinable from information given.

2. A body of mass 3.0 kg, specific gravity of 4, and weight of 29.4 N is moving in a horizontal circular path of radius 2.0 m at a constant speed of 5.0 m/sec. A rope is fastened to the body and to a post at the centre of the circle. The body

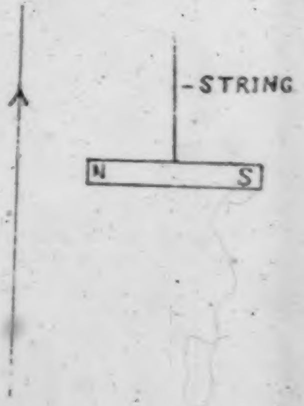
   A. undergoes constant acceleration towards the centre of the circle
   B. rotates at 1.0 rev/sec
   C. has constant momentum
   D. has steadily increasing kinetic energy
   E. does not change velocity as it moves around the circle.

3. A hydrogen ion having a charge of 1.6×10^{-19} coulombs is situated in a vacuum in an electric field. The mass of the ion is 1 a.m.u. and its atomic number is 1. As the ion moves under the influence of the electric field

   A. no work is done on the ion
   B. the kinetic energy of the ion increases
   C. the speed of the ion remains constant
   D. the ion acquires potential energy
   E. the charge on the ion slowly decreases.

4. When electrons flow upward in the wire as indicated, the N-pole of the magnet suspended near the wire as shown will experience a force which

   A. will attract it toward the wire
   B. will tend to push it "into" the page
   C. will repel it from the wire
   D. will tend to push it "out" of the page
   E. will tend to raise it.



5. Three forces have magnitudes of 2 N, 1 N, and 5 N. They may be combined so as to produce a resultant force of magnitude

   A. zero
   B. 1 N
   C. any amount between 1 N and 5 N
   D. any amount between 3 N and 4 N
   E. greater than 8 N.

6. An astronaut weighs 160 lbs. when on earth's surface. Into his space capsule he takes a set of bathroom scales to weigh himself in space. When the capsule is orbiting the earth 100 miles above the surface he stands on the bathroom scales which have been glued to the floor. He finds that the scales read zero. This is probably because

   A. at this height gravity is so small that his weight is negligible
   B. the glue holding the scales to the floor has fouled the workings of the scales
   C. there is no gravitational pull at this height
   D. he and the scales are both accelerating towards the earth at the same rate
   E. the buoyancy of the pressurized air in the satellite supports his weight.

7. In a region of space there is a magnetic field directed vertically down. A straight length of wire in an east-west orientation is moved south at a steady speed through the field. Long conducting wires hang from each end of the magnetic field. As the straight length of wire is being moved an induced electron current flows through it. The direction of the induced current is such that it tends to force the wire

   A. downward
   B. upward
   C. to the south
   D. to the north
   E. to rotate.

8. A loaded toboggan weighing 40 lbs. is pulled along by a rope with a force of 10 lbs. The force is applied in a direction 40° above the horizontal. When the toboggan has been pulled along the ground for a distance of 200 ft. at a speed of 8 ft/sec, the work done is

   A. not determinable from the information given

B. 10 X 200 X cos 40° ft lbs
C. 40 X 8 X sin 40° ft lbs
D. equal to the kinetic energy of the loaded toboggan
E. equal to the product of the momentum and the speed of the loaded toboggan.

9. An object carrying 2.0 X 10^{-4} coulombs of positive charge enters a uniform magnetic field in a direction perpendicular to the field direction. The mass of the object is 1.0 grams. Which one of the following changes would decrease the magnitude of the force experienced by the object?

A. Alter the direction of motion of the charge to parallel the field direction
B. Increase the speed of the object
C. Decrease the mass of the object
D. Change the charge on the object to 2.0 X 10^{-4} coulombs of negative charge
E. Add more positive charge to the object.

10. A raindrop has a radius of 0.1 cm and a mass of 4 milligrams. A steady wind is blowing and the raindrop, in falling to the earth, has attained a constant velocity due to the combined effects of gravity and air resistance. For the remainder of its fall

A. its momentum will continue to change
B. its kinetic energy will continue to increase
C. the resultant of all forces acting on it is zero
D. its potential energy will remain constant
E. the gravitational pull will decrease.

11. An unmagnetized iron nail can become magnetized by induction if it is placed in a magnetic field. When such a nail is placed between the poles of a powerful horseshoe magnet, the magnetic domains in the nail

A. rotate to become aligned with the magnet's field
B. grow in size if they are already aligned with the magnet's field
C. start to become magnetized as soon as the nail is placed in the magnetic field
D. move toward the places on the nail where the magnetic poles will be
E. begin to spin around and thus produce the magnetism in the nail.

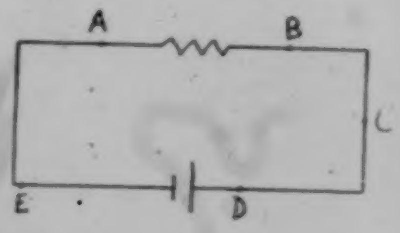
12. A battery supplies 2.0 amperes of current through a wire whose resistance is 0.1 ohms per foot. The circuit is connected for 5.0 minutes. An ammeter, a voltmeter, and a wattmeter are connected in the circuit. The number of coulombs of charge which moved through the circuit may be known

    A. by direct reading from the ammeter
    B. from the product of the wattmeter reading and the time in hours
    C. from the product of the ammeter reading and the time in seconds
    D. from the product of the current, the voltage, and the time in seconds
    E. only if more information is provided.

13. A mass of 1.0 slug is uniformly accelerated from 30 ft/sec to zero in a time of 3.0 sec. The magnitude of the force acting is

    A. 10 lbs
    B. 20 lbs
    C. not determinable from the information given
    D. 90 lbs
    E. 2.8 lbs.

14. In the circuit shown at right a cell and a resistor are connected by wires of uniform composition and cross-section. When a precise voltmeter is used to measure the potential difference between various points in the circuit, the potential difference will be



    A. the same between successive pairs of points, i.e., the same between A and B as it is between B and C, etc.
    B. greatest between A and B
    C. zero between E and A
    D. greatest between D and C
    E. greater between E and D than it is between A and D.

15. An object of weight 5.0 lbs is released from a helicopter flying horizontally at 80 mi/hr at a height of 1000 ft. As the object is falling to the ground it will have a horizontal acceleration of

    A. 80 mi/hr

B. 1000 ÷ 80 ft/mi/hr
C. 32 ft/sec^2
D. 1000 ft/min/sec
E. zero.

16. A cell with an electromotive force of 2.0 volts has an internal resistance of 0.3 ohms when it is connected with an external of 9.7 ohms. When the external resistance is reduced to 1/10 as much, the current through it will

A. remain constant
B. increase by 10 times
C. increase by more than 10 times
D. increase, but by less than 10 times
E. decrease to 1/10 as much.

17. A capacitor consisting of two widely separated conducting plates is connected in series with a battery whose e.m.f. is 5.0 volts and a wire resistor of 10 ohms resistance. As the two plates are slowly moved closer together

A. positive charge will flow through the resistor
B. electrons will flow from the negative terminal of the battery to the plate connected with it
C. the electrical energy stored in the capacitor will decrease
D. an alternating current will flow through the resistor
E. the amount of charge stored on each plate of the capacitor will decrease.

18. A metre stick, pivoted at its centre, has a mass of 400 gs suspended at the 20 cm mark, a mass of 50 gs suspended at the 70 cm mark and a mass of 60 gs suspended at the 40 cm mark. A mass of 20 gs

A. at the 90 cm mark will produce equilibrium
B. at the 10 cm mark will produce equilibrium
C. at the 60 cm mark will produce equilibrium
D. at the 40 cm mark will produce equilibrium
E. cannot produce equilibrium no matter where it is placed on the metre stick.

19. Two objects, one of mass 1 slug, the other of mass 2 slugs are released simultaneously from the top of a 100 ft high building. As they fall to the ground

A. the larger mass acquires kinetic energy at the same rate as the smaller mass
B. the momenta increase at equal rates
C. their velocities remain constant
D. the larger mass loses potential energy more rapidly than the smaller mass
E. equal forces act on the two masses.

20. The centre of gravity of a large closed book which is on a level table is at the centre of page 500. Then the book is opened at the middle. The open book resting on its back on the table will have its centre of gravity

A. lower than when it was closed
B. at the centre of page 500
C. at the centre of page 250
D. at the spine of the book but at the level of page 500
E. at the spine of the book where it touches the table.

21. A car going at 80 ft/sec slows down and stops in a distance of 200 ft. The coefficient of friction effective in stopping the car is

A. less than 0.20
B. between 0.20 and 0.45
C. greater than 0.60
D. between 0.45 and 0.60
E. not determinable from information given.

22. Electric charge is transferred by metals and by nonmetallic electrolytes in solution. Most metals and solid electrolytes are crystalline, while many nonmetals are not. As far as the transfer of electric charge is concerned, the main difference between metallic conductors and insulators is

A. metallic conductors have a greater supply of free electrons than insulators
B. metallic conductors have a greater number of negative than positive charges
C. atoms are more tightly packed in conductors than in insulators
D. both electrons and protons are free to move in conductors but only electrons in insulators
E. electrons are able to be transferred with much greater speed in insulators than in conductors.

23. 10 light bulbs are each rated at 100 watts when operated with a potential difference of 100 volts. In a circuit whose potential difference is 100 volts the bulbs are connected, all in parallel. In another circuit whose potential difference is 100 volts all the bulbs are connected in series. Each circuit is operated for 1 hour. The electrical energy used is

A. the same in each circuit
B. about 100 times greater in the parallel circuit
C. about 10 times greater in the parallel circuit
D. about 100 times greater in the series circuit
E. about 10 times greater in the series circuit.

24. Two magnetic poles which are (c) cm apart experience a force of (D) dynes. If the two poles are moved to such a separation that the force is (2D) dynes, then their distance apart must be

A. (2c) cm
B. $\left(\frac{c}{\sqrt{2}}\right)$ cm
C. (c) cm
D. (½c) cm
E. none of the above distances.

25. An atomic nucleus has a positive charge and a linear size of less than 10^{-14} cm. The binding energy of the nucleus is

A. the energy which would be required to separate it into its constituent protons and neutrons
B. usually about equal to the energy needed to remove an outer electron from its orbit
C. greatest for an ordinary hydrogen nucleus
D. released when an atom disintegrates into its separate neutrons and protons
E. the same for all isotopes of an element.

26. A small laboratory cart is released from rest at the top of a long inclined plane. During the first second it travels 10 inches along the plane. During the next second it will go

A. 10 inches
B. 20 inches
C. 25 inches
D. 30 inches
E. more than 30 inches.

27. An iron ball has a density 9 times as great, and a mass 3 times as great, as that of a wooden ball. The two balls, initially at rest, are acted upon by equal horizontal forces for a period of 3.0 sec. At the end of 2.0 sec of this time

A. the wooden ball will have the same momentum as the iron one
B. the two balls will have equal kinetic energies
C. the two balls will have equal velocities
D. the two balls will have equal accelerations
E. the two balls will be worked on with equal power.

28. The radiation in the Van Allen belts consists mainly of high speed protons and electrons. The protons are nearly 2000 times as massive as the electrons although protons and electrons have equal amounts of charge. The Van Allen belts do not extend down to the surface of the earth because

A. the earth's magnetic field is not strong enough at the surface to hold them there
B. the protons and electrons neutralize each other
C. the sun's heat is not strong enough near the earth's surface to ionize air molecules
D. the moving charged particles collide with air molecules in denser air near the earth's surface thus reducing their speeds and the force on them due to the earth's magnetic field
E. solar winds blow them away.

29. Work of 49 joules is done in lifting a mass of 0.5 kg 10 m from point A to point B in a gravitational field. The time required to do the work is 20 sec. If gravitational potential difference, similar to electrical potential difference, was calculated ~~for~~ A and B, it would be

A. about 10 N/kg
B. about 98 J/kg
C. about 49 J
D. about 2.5 J/sec
E. about 4.9 J/m.

30. A wire 8.0m long carries a current of 5.0 amperes through a magnetic field of 1.0×10^{-2} telsa. The wire is perpendicular to the field direction. The motor-effect force on the wire is

A. 4.0×10^{-1} N

B. 1.3×10^{-2} N

C. 130 N

D. 0.63×10^{-1} N

E. none of the above.

31. A 2.0 watt resistor is to be employed with a 6.0 volt potential difference. The resistance of the resistor is

A. 3.0 ohms
B. 0.33 ohms
C. 18 ohms
D. 2.0 ohms
E. 0.17 ohms.

32. A steel ball of mass 10 gs is released from a height of 4.0 m above the ground. It falls to the ground in about 0.9 sec and attains a speed of about 9 m/sec just before hitting the ground. The weight of the ball is about 0.1 N. During the ‸ that the ball is falling
time

A. the momentum of the ball is constant
B. the acting force on the ball is counter-balanced by a reacting force on the earth
C. the terminal speed of the ball in flight is reached
D. the acting force is gravity, and the reacting force is the inertia of the ball
E. Newton's third law does not apply to this kind of situation.

33. A 4 foot length of wire having a resistance of 10 ohms is connected to a battery whose e.m.f. is 2.0 volts. As an electron moves along the wire

A. no work is done on the electron
B. the kinetic energy of the electron increases
C. the speed of the electron remains steady
D. the electron acquires potential energy
E. the work done on the electron is lost as heat through collisions.

34. A mass of 2.0 kg is accelerated upwards at 2.0 m/sec^2 by means of a force applied through a rope. The total force exerted by the rope is

A. 15.6 N
B. greater than 25 N
C. 4.0 N
D. 23.6 N
E. 19.6 N

35. When at rest M_1 = kg and M_2 = 2kg. The two masses may become equal

A. under no conditions
B. provided that the speed of M_1 is 200 m/sec and M_2 is 100 m/sec
C. provided that the speed of M_1 is equal to the speed of light and M_2 is equal to one-half the speed of light
D. provided M_2 is at rest and M_1 has a speed between one-half the speed of light and the speed of light
E. if M_2 were located at a point sufficiently far out in space.

36. A test charge of 4.0×10^{-5} coulombs is situated 20 cm from a charged ball. The test charge experiences a repelling force of 2.0×10^{-4} N due to the electric field of the charged ball. The magnitude of the electric field intensity at the test charge is

A. $4.0 \times 10^{-5} \div 2.0 \times 10^{-4}$ coulombs/N
B. $4.0 \times 10^{-5} \times 2.0 \times 10^{-4}$ N-coulombs
C. $2.0 \times 10^{-4} \div 400$ N/cm^2
D. $4.0 \times 10^{-5} \div 400$ coulombs/cm^2
E. $2.0 \times 10^{-4} \div 4.0 \times 10^{-5}$ N/coulomb.

37. An object which weighs 10 lbs on the earth's surface is raised to an altitude of 100 miles above the earth. The pull of gravity on the object at this height would be

A. between 9 and 10 lbs
B. about 5 lbs
C. zero
D. between 10 and 11 lbs
E. greater than 11 lbs.

38. An ammeter for measuring currents of up to 5.0 amperes has a moving coil of 100 turns located between the poles of a horseshoe magnet. Because the coil is easily melted and cannot carry more than 1.0 x 10^{-3} amperes without damage

A. the ammeter must always be connected in parallel across a circuit branch
B. a high resistance is connected in series with the coil
C. a special diode is connected to prevent currents greater than 1.0 x 10^{-3} amperes from passing through the meter
D. a shunt or by-pass is provided in the instrument so that only a small fraction of the current being measured will go through the coil
E. this ammeter is only suitable for measuring very small currents less than 1.0 x 10^{-3} amperes.

39. Two points in an electric circuit have a potential difference of 3.0 volts. The work done transferring 12 coulombs of charge between these two points is

A. 0
B. 4 joules
C. 0.25 joules
D. 36 joules
E. not determinable from the information given.

40. A bucket of weight 2,000 N can hold 40,000 N of concrete. The bucket is operated by a motor whose maximum power output is 3,000,000 watts. The empty bucket is to be raised slowly but at a constant speed. The power expenditure in raising the bucket is

A. 2000 x 6 watts
B. 42000 x 6 watts
C. 42000 ÷ 6 watts
D. not determinable from the information given
E. greater than 3,000,000 watts.

APPENDIX C

The Discrepancy Detection Test

DISCREPANCY DETECTION TEST

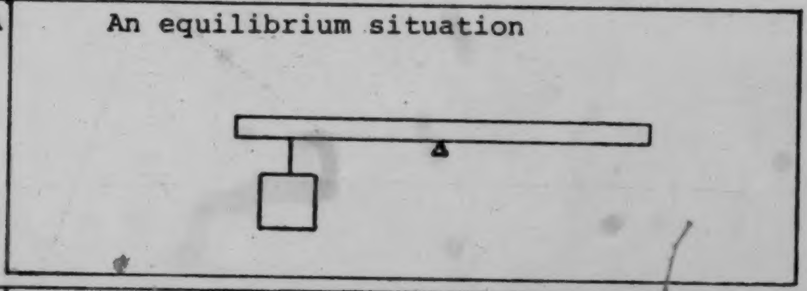
The following test consists of a series of diagrams, often with a few words of description. Some diagrams contain contradictions or discrepancies between what is shown and what would be expected from a knowledge of physics, while other diagrams do not contain such contradictions.

Examine each diagram carefully. If you think that there is no contradiction between what is shown and the principles of physics, write the letter "C", for correct, beside that diagram's number on the answer sheet. If you think that there is some contradiction, state both what would be expected in that situation from physics principles, and also what seems to be wrong with what is observed.

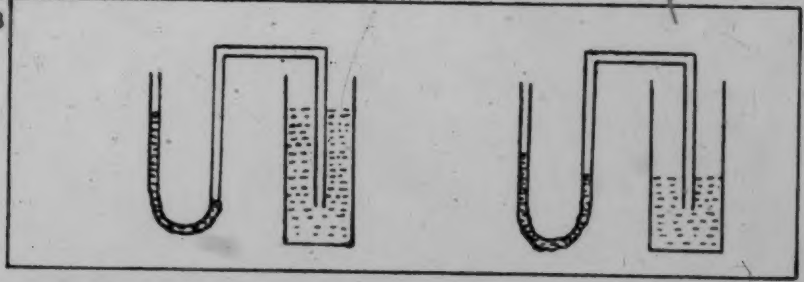
In all examples assume that there is no friction or air resistance unless the item specifically states that friction is a factor.

Examples of items:

A



B



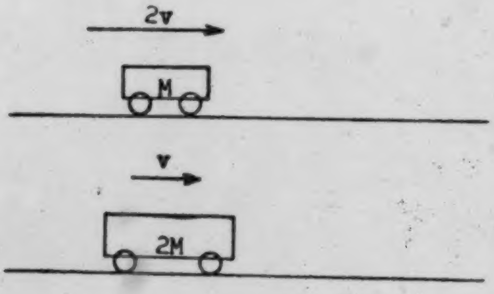
Sample Answers:

A. For equilibrium clockwise moments equal counter clockwise moments. No forces produce clockwise moments in this figure, so there is no equilibrium.

B. C.

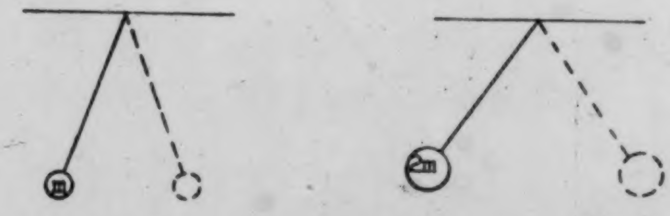
1.

These two masses have equal kinetic energies



2.

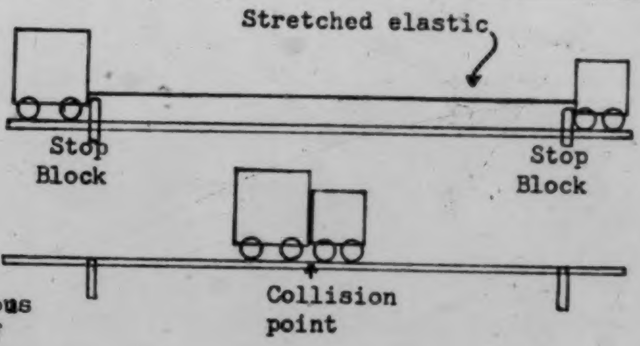
These pendula vibrate with equal frequency



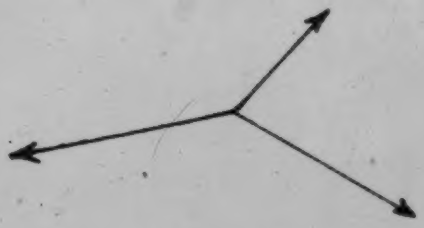
3.

Before:

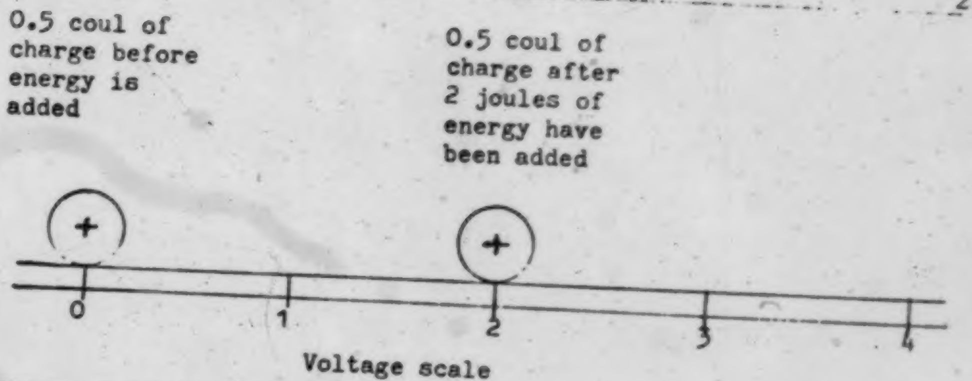
After simultaneous release of stop blocks:


Stretched elastic
Stop Block
Stop Block
Collision point

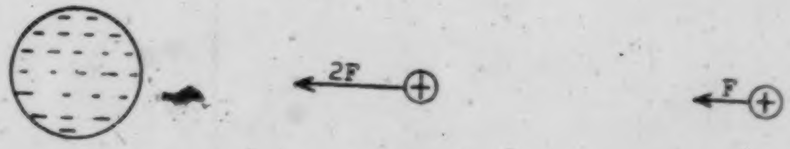
4. Three forces in equilibrium



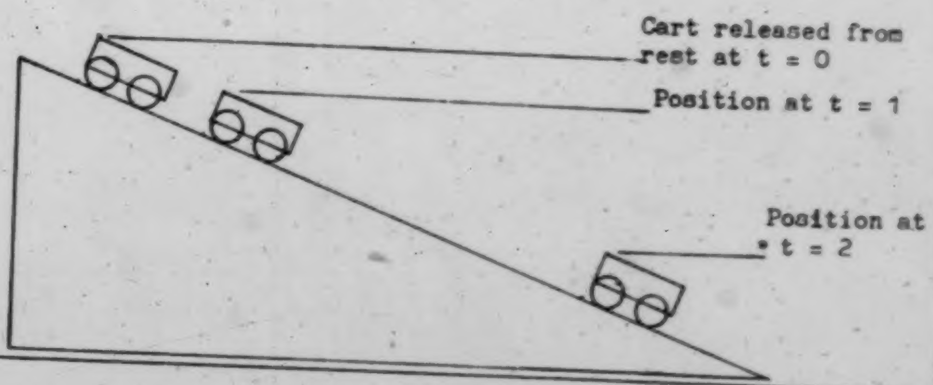
5.


0.5 coul of charge before energy is added
0.5 coul of charge after 2 joules of energy have been added
+
+
0
1
2
3
4
Voltage scale

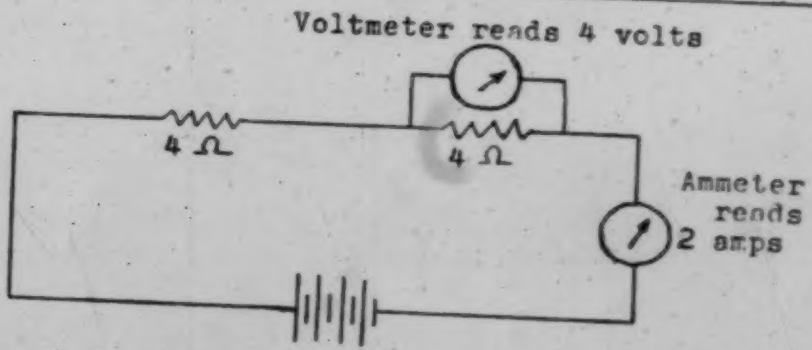
6.


2F
F

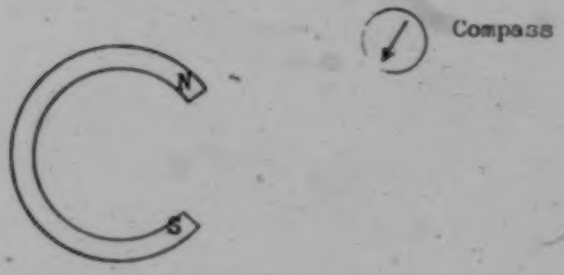
7.


Cart released from rest at t = 0
Position at t = 1
Position at t = 2

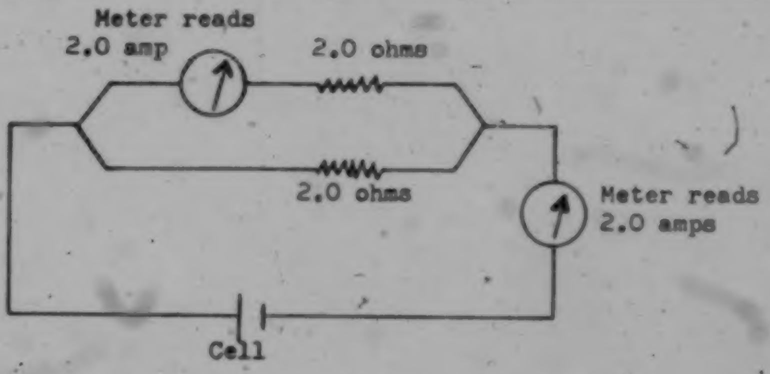
8.


Voltmeter reads 4 volts
4 Ω
4 Ω
Ammeter reads 2 amps

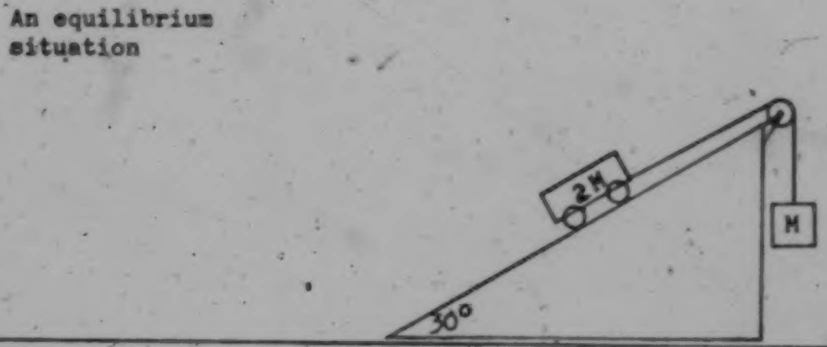
9.


Compass
N
S

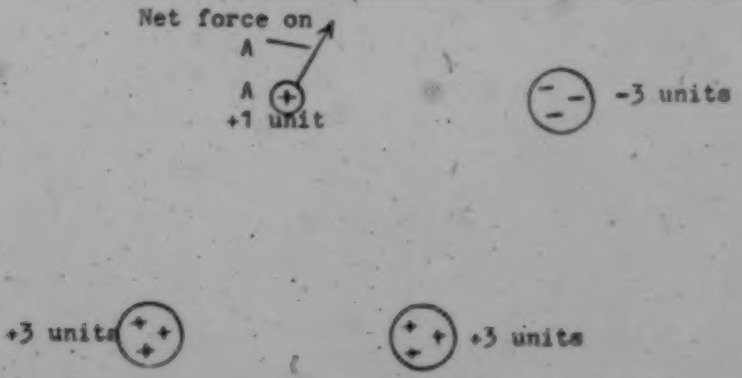
10.


Meter reads
2.0 amp
2.0 ohms
2.0 ohms
Meter reads
2.0 amps
Cell

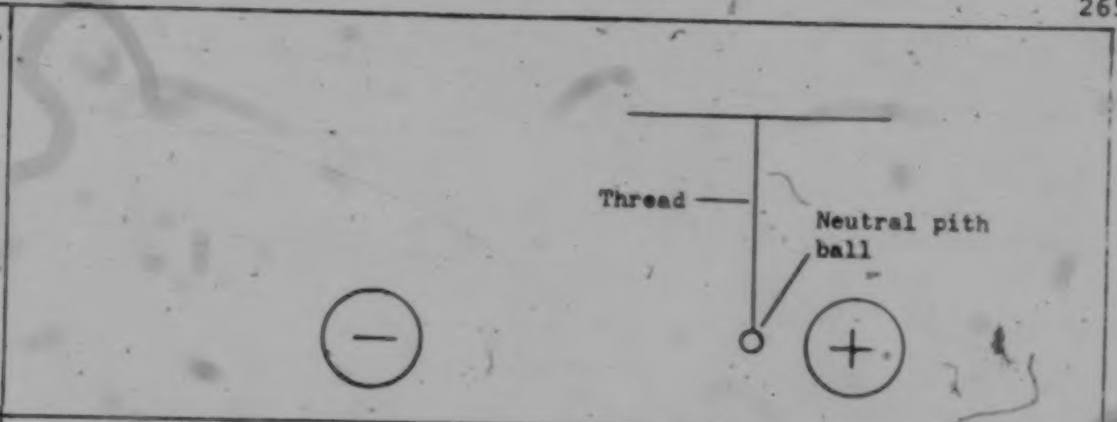
11.


An equilibrium
situation
2M
M
30°

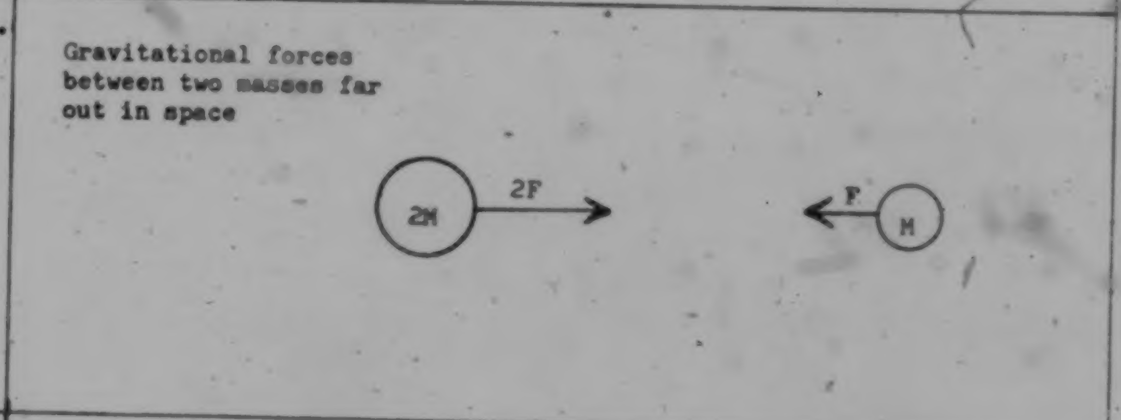
12.


Net force on
A
A
+1 unit
-3 units
+3 units
+3 units

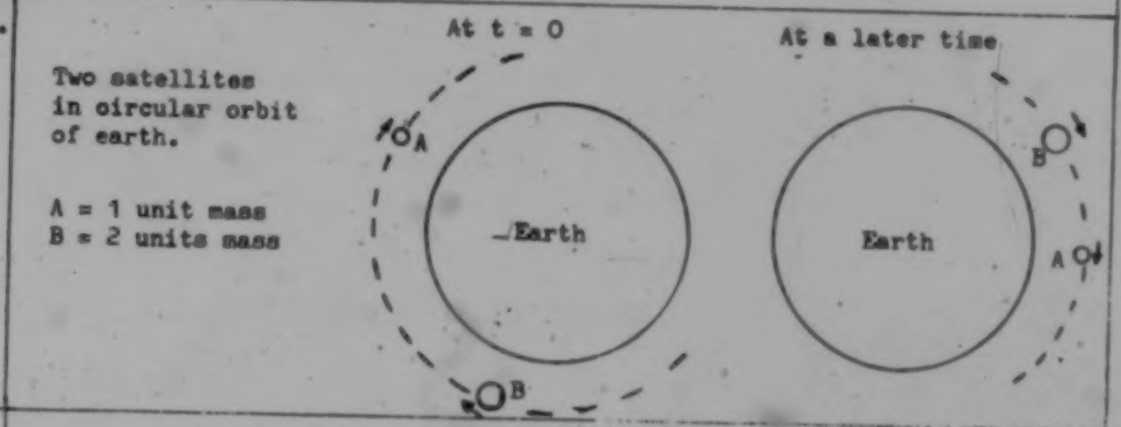
13.


Thread
Neutral pith ball
−
+

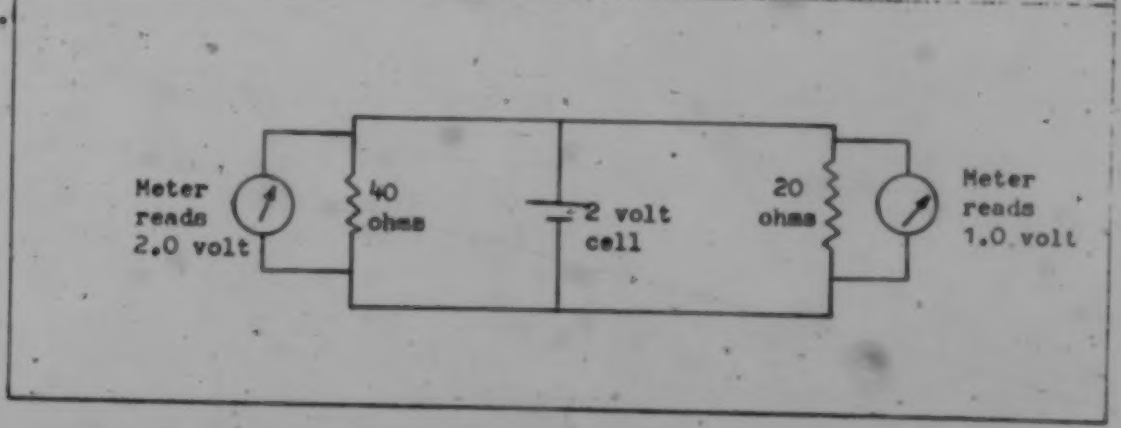
14.


Gravitational forces between two masses far out in space
2M
2F
F
M

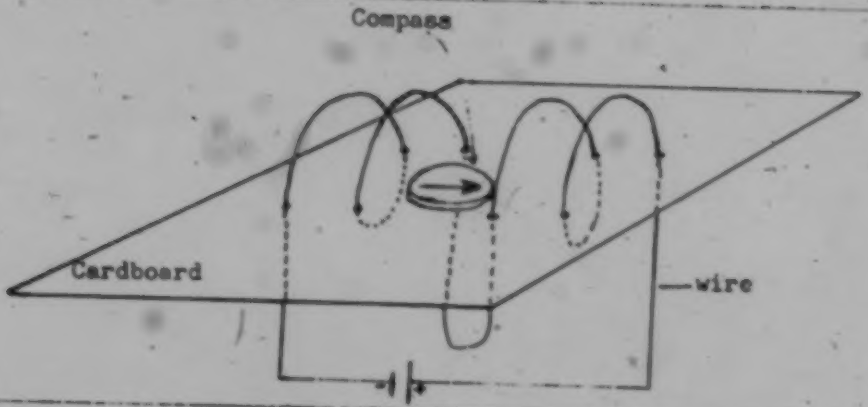
15.


At t = 0
At a later time
Two satellites in circular orbit of earth.
A = 1 unit mass
B = 2 units mass
A
B
Earth
Earth
B
A

16.

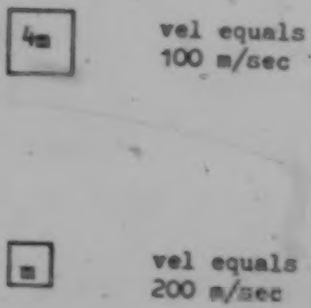

Meter reads 2.0 volt
40 ohms
2 volt cell
20 ohms
Meter reads 1.0 volt

17.

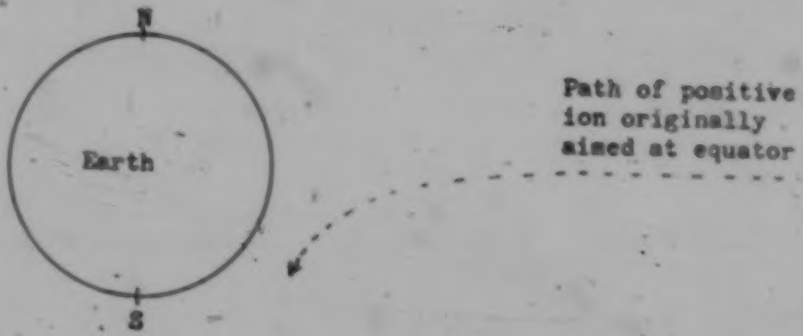

Compass
Cardboard
wire

18.

The mass-energies of the two objects at right are equal


4m
vel equals 100 m/sec
m
vel equals 200 m/sec

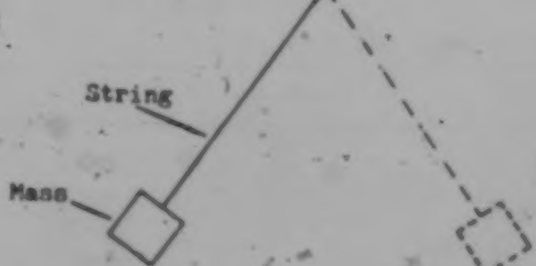
19.



Path of positive ion originally aimed at equator

20.

Work is done by the string during each oscillation of the mass


String
Mass

APPENDIX D

Raw Scores

RAW SCORES

Listed below are the raw scores for the students of the study. The test means have been substituted for missing data as described in Chapter 5. The following abbreviations have been used for the column heads:

ID - identification number

HCT - High Context Test

LCT - Low Context Test

DDT - Discrepancy Detection Test

SCAT-V - Cooperative School and College Ability Test - Verbal

SCAT-Q - Cooperative School and College Ability Test - Quantitative

HFT - Hidden Figures Test

CWS - Category Width Scale

P30 - Physics 30 Achievement Test

RAW SCORES (continued)

ID	HCT	LCT	DDT	SCAT-V	SCAT-Q	HFT	CWS	P30
1	9	9	8	51	26	16	30	39
2	7	8	7	51	37	14	44	40
3	8	11	10	54	42	12	26	40
4	8	10	8	44	36	9	32	50
5	8	13	9	48	41	12	30	51
6	15	13	13	44	36	18	33	50
7	6	5	8	38	42	11	30	34
8	6	12	3	48	39	9	46	40
9	15	18	13	51	45	20	30	60
10	6	10	10	44	36	11	30	47
11	13	13	12	48	48	24	37	55
12	5	11	9	44	36	13	28	43
13	9	7	11	54	44	12	47	52
14	9	5	9	44	36	24	26	51
15	5	11	8	53	35	15	25	47
16	5	9	7	53	43	11	21	52
17	10	11	10	56	39	14	38	53
18	10	6	8	59	33	10	36	38
19	8	10	5	34	20	15	20	30
20	6	11	8	44	36	15	21	40
21	6	12	7	38	37	13	35	39
22	12	10	13	56	43	8	45	61
23	11	13	13	44	32	11	35	48
24	9	14	12	55	32	17	40	55
25	8	5	5	34	28	9	30	11
26	10	8	6	44	36	14	34	28
27	6	5	7	44	36	14	42	26
28	7	17	10	55	36	8	28	37
29	7	10	6	44	36	11	23	48
30	8	12	8	55	41	11	30	43
31	5	6	7	43	39	11	39	26
32	10	13	9	46	32	16	45	45
33	8	10	8	46	41	11	30	38
34	9	14	8	41	35	9	25	38
35	13	12	10	44	36	21	36	49

RAW SCORES (continued)

ID	HCT	LCT	DDT	SCAT-V	SCAT-Q	HFT	CWS	P30
36	3	8	10	33	16	7	31	13
37	14	13	12	44	36	17	35	53
38	6	7	10	49	37	9	15	38
39	9	16	12	48	33	13	44	59
40	6	7	8	37	31	11	28	34
41	9	15	6	49	36	6	23	46
42	6	10	10	45	31	11	31	38
43	16	13	14	44	36	15	25	62
44	8	10	9	42	45	5	30	40
45	10	16	10	51	34	13	29	46
46	10	10	9	46	46	13	33	59
47	7	13	7	46	41	8	33	41
48	6	11	2	44	34	6	26	36
49	5	10	8	59	42	11	30	34
50	7	13	14	41	39	11	23	43
51	5	9	7	41	42	2	38	34
52	8	8	8	55	45	11	30	30
53	9	7	8	37	42	11	30	35
54	6	9	8	50	39	11	30	35
55	7	6	6	42	34	19	13	35
56	10	13	7	55	47	15	31	57
57	7	8	5	50	38	9	22	24
58	5	7	6	45	42	3	45	37
59	6	9	7	51	37	1	26	35
60	7	11	7	46	34	19	30	44
61	11	12	13	52	34	24	12	56
62	9	12	14	43	43	9	36	49
63	9	12	8	56	45	11	30	53
64	8	11	3	35	40	10	10	29
65	4	8	4	21	24	21	29	27
66	12	9	7	52	43	10	20	38
67	10	12	8	57	43	11	30	52
68	7	6	8	39	31	11	30	28
69	11	10	7	53	43	23	40	40
70	9	6	4	53	37	8	34	30

RAW SCORES (continued)

ID	HCT	LCT	DDT	SCAT-V	SCAT-Q	HFT	CWS	P30
71	9	14	8	57	43	11	30	56
72	7	6	8	47	41	11	30	26
73	8	6	4	44	36	6	18	30
74	5	9	8	55	43	11	30	33
75	7	10	10	53	38	7	26	44
76	7	10	14	56	38	3	21	46
77	2	3	3	44	36	14	23	19
78	6	7	9	36	31	11	30	26
79	4	10	7	46	28	11	30	19
80	7	11	9	52	40	12	40	28
81	7	9	6	43	39	4	43	35
82	6	12	4	55	32	23	52	28
83	9	14	8	47	30	9	25	34
84	7	6	8	38	28	4	17	30
85	8	10	7	47	34	10	33	28
86	14	16	13	36	37	11	37	55
87	6	11	7	37	23	18	30	28
88	7	12	9	44	36	5	30	31
89	6	3	4	33	32	7	23	19
90	7	9	6	44	36	14	24	39
91	2	8	6	45	33	14	19	32
92	4	9	3	49	35	8	25	25
93	7	7	8	31	38	6	33	23
94	3	3	5	34	35	11	33	33
95	7	12	7	36	33	12	34	29
96	7	6	6	38	42	6	46	37
97	6	7	9	42	38	11	30	25
98	4	8	5	46	33	6	37	26
99	7	9	8	38	28	5	9	41
100	9	10	8	53	28	6	29	41
101	5	5	9	44	36	1	33	40
102	8	8	4	40	27	13	24	41
103	7	7	8	11	26	1	17	23
104	10	12	7	40	46	17	36	47
105	8	10	9	40	30	14	38	18

RAW SCORES (continued)

ID	HCT	LCT	DDT	SCAT-V	SCAT-Q	HFT	CWS	P30
106	9	11	8	37	39	17	27	30
107	7	13	9	44	36	1	15	45
108	4	11	5	50	42	16	16	34
109	6	8	4	44	36	1	34	26
110	7	11	5	45	32	14	31	35
111	8	9	9	18	18	2	26	41
112	1	3	2	57	38	5	37	20
113	8	12	7	54	37	11	30	31
114	8	9	7	44	38	11	30	27
115	7	5	7	36	24	11	30	24
116	3	2	3	44	36	1	58	16
117	5	6	7	44	36	16	20	21
118	3	10	5	21	31	9	16	28
119	5	13	9	44	36	2	28	42
120	3	6	5	57	42	17	28	25
121	6	8	6	31	25	8	8	31
122	8	8	6	44	36	13	27	33
123	3	5	1	44	36	3	20	29
124	7	6	9	34	39	7	25	25
125	6	6	3	49	32	5	26	22
126	6	9	6	34	33	3	28	27
127	6	8	1	44	36	15	26	40
128	7	8	6	37	39	12	8	22
129	8	12	11	25	37	15	37	46
130	7	11	7	36	40	11	31	36
131	7	6	7	12	11	11	30	20
132	8	10	4	44	36	11	33	29
133	8	7	4	44	36	7	19	37
134	7	10	8	32	39	14	30	33
135	14	14	7	44	36	10	26	52
136	6	8	5	44	36	12	36	33
137	7	10	5	48	42	12	26	35
138	7	10	6	47	32	5	37	20
139	7	10	5	35	17	13	19	17

APPENDIX E

Additional Tables

Table 13

Normalized Weights Applied to the Criterion and Predictor Variables for the First and Second Canonical Correlations

Criterion Variable		Predictor Variable	
First Canonical Correlation Coefficient = 0.76			
High Context Test	.66	SCAT-Verbal	-.06
Low Context Test	.57	SCAT-Quantitative	-.13
Discrepancy Detection Test	.50	Physics 30 Test	.98
		Hidden Figures Test	.15
		Category Width Scale	.07
Second Canonical Correlation Coefficient = 0.19			
High Context Test	.61	SCAT-Verbal	-.77
Low Context Test	-.78	SCAT-Quantitative	.43
Discrepancy Detection Test	.13	Physics 30 Test	-.10
		Hidden Figures Test	.32
		Category Width Scale	.33

APPENDIX F

The Category Width Scale

PETTIGREW ESTIMATION SCALE

In this questionnaire each item gives an average value of some quantity. Then, four possible maximum values of the quantity are presented, and also four possible minimum values. You are asked to estimate the maximum value which might be observed, and also the minimum value, by selecting from the alternatives available.

Sample item:

1. In July the average number of cars per day crossing a certain city bridge is 8000. What do you think:

   a) is the largest number of cars crossing in a single day in July?

      1. 200,000 ( ) 3. 50,000 ( )
      2. 100,000 ( ) 4. 15,000 ( )

   b) is the smallest number of cars crossing in a single day in July?

      1. 100 ( ) 3. 1000 ( )
      2. 500 ( ) 4. 5000 ( )

Consider each statement carefully before making your decision.

Place a check mark (✓) in the bracket beside the maximum value you select in (a), and beside the minimum value you select in (b).

1. It has been estimated that the average width of windows is *34 inches*. What do you think:

   a. is the width of the widest window...

   1. 1,363 inches ( ) 3. 48 inches ( )
   2. 341 inches ( ) 4. 81 inches ( )

   b. is the width of the narrowest window...

   1. 3 inches ( ) 3. 11 inches ( )
   2. 18 inches ( ) 4. 1 inch ( )

2. Ornithologists tell us that the best guess of the average speed of birds in flight would be about *17 m.p.h.* What do you think:

   a. is the speed in flight of the fastest bird...

   1. 25 m.p.h. ( ) 3. 73 m.p.h. ( )
   2. 105 m.p.h. ( ) 4. 34 m.p.h. ( )

   b. is the speed in flight of the slowest bird...

   1. 10 m.p.h. ( ) 3. 12 m.p.h. ( )
   2. 2 m.p.h. ( ) 4. 5 m.p.h. ( )

3. Weather officials report that during this century Washington, D.C., has received an average rainfall of *41.1 inches* annually. What do you think:

   a. is the largest amount of rain that Washington has received in a single year during this century...

   1. 82.4 inches ( ) 3. 63.7 inches ( )
   2. 45.8 inches ( ) 4. 51.2 inches ( )

   b. is the smallest amount of rain that Washington has received in a single year during this century...

   1. 20.2 inches ( ) 3. 9.9 inches ( )
   2. 36.3 inches ( ) 4. 29.7 inches ( )

4. An average of *58 ships* entered or left New York harbor daily during the period from 1950 through 1955. What do you think:

a. was the largest number of ships to enter or leave New York in a single day during this period...

1. 69 ships ( )
2. 153 ships ( )
3. 76 ships ( )
4. 102 ships ( )

b. was the smallest number of ships to enter or leave New York in a single day during this period...

1. 34 ships ( )
2. 3 ships ( )
3. 16 ships ( )
4. 43 ships ( )

5. Boating experts estimate that the average speed of all sailing crafts in America is around *4.1 knots*. What do you think:

a. is the speed of the fastest sailing boat in America...

1. 8.2 knots ( )
2. 30.7 knots ( )
3. 5.9 knots ( )
4. 21.3 knots ( )

b. is the speed of the slowest sailing boat in America...

1. 3.3 knots ( )
2. 0.6 knots ( )
3. 2.2 knots ( )
4. 1.2 knots ( )

6. When all of the world's written languages are considered, linguists tell us that the average number of verbs per language must be somewhere around *15,000*. What do you think:

a. is the largest number of verbs in any single language...

1. 21,000 ( )
2. 18,000 ( )
3. 50,000 ( )
4. 30,000 ( )

b. is the smallest number of verbs in any single language...

1. 1,000 ( )
2. 13,000 ( )
3. 5,000 ( )
4. 10,000 ( )

7. The average muzzle to tail length of a sample of 1,000 German Shepherd dogs is *40.3 in.* What do you think:

   a. is the length of the longest Shepherd dog in the sample...

      1. 60.4 inches ( )
      2. 47.8 inches ( )
      3. 44.1 inches ( )
      4. 54.2 inches ( )

   b. is the length of the shortest Shepherd dog in the sample...

      1. 34.6 inches ( )
      2. 28.4 inches ( )
      3. 19.7 inches ( )
      4. 36.9 inches ( )

8. The average population of South America countries is approximately *8.6 million* people each. What do you think:

   a. is the population of the most populated country in South America...

      1. 11.2 million ( )
      2. 54.7 million ( )
      3. 23.6 million ( )
      4. 129.1 million ( )

   b. is the population of the least populated country in South America...

      1. 7,000 ( )
      2. 6.2 million ( )
      3. 2.4 million ( )
      4. 29,000 ( )

9. A Stanford University home economist has estimated the average American spends *55 minutes* of his day eating. What do you think:

   a. is the longest eating time of any single American...

      1. 185 minutes ( )
      2. 125 minutes ( )
      3. 245 minutes ( )
      4. 90 minutes ( )

   b. is the shortest eating time of any single American...

      1. 16 minutes ( )
      2. 4 minutes ( )
      3. 38 minutes ( )
      4. 27 minutes ( )

10. The average number of churches per religious denomination in the United States is estimated to be *511*. What do you think:

a. is the largest number of churches of a single religious denomination in the U.S.A....

1. 4,833 ( )
2. 757 ( )
3. 1,219 ( )
4. 39,801 ( )

b. is the smallest number of churches of a single religious denomination in the U.S.A....

1. 313 ( )
2. 146 ( )
3. 1 ( )
4. 23 ( )